DEVELOPMENTAL PSYCHOLOGY | RESEARCH ARTICLE

Language can obscure as well as facilitate apparent-Theory of mind performance: part 1 - An exploratory study with 4 year-Olds using the element of surprise

Barlow C Wright

Abstract: Language is integral to children's Theory of Mind (ToM) development. Here, we also considered whether language emerges as important because tasks assess ToM through language. Fifty-five typically-developing 4-year-olds completed eight false-belief tasks via film clips, responding verbally or by pointing, plus explaining each response. Each clip was then played out to its conclusion and children’s surprise to expected versus unexpected searches and outcomes was measured. Total performance using surprise was 71% higher than the standard index and 3-times as high as verbally explained ToM. Contrasts between four intersections of likely/unlikely searches with plausible/implausible object-retrievals, revealed children were most surprised when both search and retrieval were unlikely-implausible. Contrastingly, surprise for unlikely-search/plausible-retrieval, was the only sub-task predicting variation in verbally explained ToM. For total scores, gender but not surprise, predicted verbally explained ToM. Indexes using surprise suggest 4-year-olds have high ToM compared to indexes heavily reliant on language. Previous findings that girls’ ToM is higher than in boys, may also stem from a reliance on language. Also, children’s ToM is more evident in pretend-contexts than in real-life-contexts. We interpret our findings as evidence that testing of ToM using low-language tasks alongside language-laden tasks may permit a more complete picture of ToM development.

Subjects: Education - Social Sciences; General Psychology; Developmental Psychology; Educational Psychology

Keywords: children; false-belief; implicit-explicit continuum; language; surprise; theory of mind

The ability routinely to assign mental states to other people (Theory of Mind—ToM) helps us bridge gaps between our own perception of someone’s action, and what actually caused those actions (i.e., people's subjective perceptions of belief states). As such, ToM includes assignment of one's own and other people's past, present, or even future beliefs; appreciation of accompanying mental states and felt emotions; present or past desires, subjective knowledge, motivations, and intentions (Premack & Woodruff, 1978; Wellman et al., 2011).

ToM has uses from social interactions, through to gaining advantage by lying or achieving the upper hand when competing for resources or even just a date, to considering altruistic acts (Guise et al., 2007; Talwar et al., 2017; Tankersley et al., 2007). Its importance in humans is further
evidenced by the fact that disruption of ToM or having an overactive ToM, can contribute to debilitating clinical conditions such as autism, bipolar disorder, or schizophrenia (Caputi & Schoenborn, 2018; Knutsen et al., 2017; Martin et al., 2014).

In terms of children’s developing ToM, “The most crucial development occurs around age 4, when they realise that thoughts in the mind may not be true.” (Astington, 1998, p. 45). This finding was first demonstrated by Wimmer and Perner (1983). In this seminal study, Wimmer and Perner introduced a task based around a child showing that s/he can hold a true belief about the location of an object in mind, whilst also demonstrating that s/he can predict what another person (actually a protagonist in a story) would do. The protagonist’s belief was previously true but is now out of date because the object was moved by another story character unbeknown to the protagonist. Whereas 4–5 year-olds readily demonstrated knowledge both of their own true belief and of the protagonist’s false-belief, 3- and early 4 year-olds seemed quite unaware about the protagonist’s false-belief.

Guajardo and Turley-Ames (2004) gave 3 to 5 year-olds a number of counterfactual and false-belief ToM tasks. Compared to a chance level of 50%, the average ToM performance for 3-year-olds was 16.2%, for 4 year-olds this was 52.3% and for 5 year-olds it was 69%. The ToM components of two longitudinal studies also support the notion of some kind of shift from 3 to 4 years (25.0% v 55.0%—Lockl & Schneider, 2007) and also for 4 versus 5 year-olds (40.8% v 66.2%—McAlister & Peterson, 2007). Bernstein et al. (2007) found very similar levels to these, across four diverse ToM tasks given as part of a large battery of reasoning tasks (3 years = 17.3%, 4 years = 57.5%, 5 years = 63.5%).

Cross-sectional findings on ToM in children find support from longitudinal studies (Wellman et al., 2011). Additionally, both types of design tend to reveal a similar progression in a number of different groups differing by culture or disability. These tasks still tend to inquire about, or measure, false-beliefs (e.g., see White et al., 2009 strange stories task; Rutherford’s stories task, Rutherford, 2004). It may partly be for these reasons that the false-belief task, although not by any means the only way to assess ToM, is often taken to remain the gold standard test, particularly when one seeks to assess ToM in children (Beaudoin et al., 2020; Scott & Roby, 2015).

ToM does seem to become more advanced after middle childhood (Caputi & Schoenborn, 2018). That said, basic ToM competences may be held much earlier than even 3 years (e.g., at 19 months or even 15 months—Onishi & Baillargeon, 2005; Repacholi & Gopnik, 1997). However, those findings are sometimes difficult to replicate (e.g., contrast Southgate et al., 2007; Yott & Poulin-Dubois, 2016 v; Kulke et al., 2018; Poulin-Dubois et al., 2018). A caveat to such contrasting findings is that failed replications are not necessarily of themselves a disproof (Baillargeon et al., 2018).

On positive ToM-like findings in infancy (e.g., Onishi & Baillargeon, 2005; Southgate et al., 2007), it is important to acknowledge that this tends not to be claimed to be a ToM competence equal to that of 4 year-olds. For example, Onishi and Baillargeon (2005, p. 257) readily concede that they may have demonstrated only “a rudimentary and implicit” ToM in infancy (for definitions see, Perner & Clements, 2000; see also, Repacholi & Gopnik, 1997; Southgate et al., 2007 for similar acknowledgement). By implicit ToM, we mean a ToM competence that the reasoner may not be aware of consciously, but which may be illustrated indirectly via indexes, such as looking times or reaction times—RT (Low & Perner, 2012). It is also worth noting that Repacholi and Gopnik’s study was on desires whereas Onishi and Baillargeon’s study seemed to test a more ToM-like ability. The difference in age of occurrence might be due to desires not being as ToM-like as is looking behaviour. However, on that view, the desires task should be solved at an earlier age than the more ToM-like looking task; but the actual findings seem the opposite of this profile.

Notwithstanding this matter, studies of implicit versus explicit ToM in children should consider the possibility that implicit ToM arrives earlier (“i.e., precedes”) and “contributes” to explicit ToM (Low, 2010, p. 598). By contrast, in typically developing adults, the mentalizing involved in explicit
ToM seems no different to that in implicit ToM (Nijhof et al., 2016). This latter conclusion is paralleled by findings where fMRI was used to identify neural sites active in implicit versus explicit ToM in adults (Naughtin et al., 2017). So, ToM may well appear on the implicit plane by 2 years or even earlier (Wiesmann et al., 2017); and then gradually become more explicit after around 3 years (Kloo et al., 2020).

An alternative way of capturing the implicit-explicit distinction and the developmental progression from one to the other, is to argue ToM is not a unitary concept; and therefore it develops in a piecemeal, fragmented or context-bound manner (Baillargeon et al., 2018). On this postulate, there is no need to limit ToM to only two spheres—implicit versus explicit. Rather, there may be any number of ToM-like abilities and it is the coming together of these into a singular concept that is marked by the passing on standard false-belief tasks. Expanding on such a view, one could argue that what occurs at around 4 years of age, may not be so much the coming on line of some ToM module (Leslie, 1994), but rather may simply be the unifying of the various ToM-like understandings into one adult-like concept of ToM.

Another important issue regarding ToM is its relationship to language. Bloom and German (2000) argue that language-related phenomena, rather than a genuinely under-developed ToM or even ToM that has thus far only appeared on the implicit plane, might be what lays behind the failures on false-belief tasks at 3 years (Baillargeon et al., 2010; Durrleman-Tame et al., 2017; Hale & Tager-Flusberg, 2003; Lohmann et al., 2005; De Mulder et al., 2019; San Juan & Astington, 2017; Wang & Su, 2009). This view can be aptly summarised as language facilitating ToM.

Certainly, the young ToM reasoner needs to interpret the situation as a social narrative and be able to explain what is happening (Lecciso et al., 2016). This implicates not only the need for an adequate vocabulary but also syntactic constructions that can efficiently represent the situation symbolically (Miller, 2006; Moran, 2013; Simmons & Singleton, 2000; Wiesmann et al., 2017). It may even be that, additional to facilitating ToM, language represents the bridge between findings on cognitive versus social processes at work in children’s ToM development (Wright & Mahfoud, 2014).

Girls tend to be more linguistically precocious/conversational than boys from mid-infancy (Hughes et al., 1999). A language facilitation view of ToM development would predict that, in line with this language advantage, girls ToM also develops earlier than that of boys (Walker, 2005). Charman et al. (2002) carried out a meta-analyses of two large datasets, concluding that girls do have superior ToM compared to boys. This was particularly evident at 3 to 4 years, which is typically, before ToM is said to have developed to above chance performance Laranjo et al., 2010; Longobardi et al., 2017; Melinder et al., 2006; Wright & Mahfoud, 2014).

However, although a few recent studies with adults do also report a gender difference in favour of girls (Wright & Wright, 2021), which conclusion is not necessarily safe. For example, Hughes et al. (1999) used false-belief tasks with 4 year-olds plus self-report measures about parental style given to the children’s parents. They did identify quite clear differences in parental perceptions about girls and boys, and differences in treatment of girls compared to boys. But there were no differences between girls’ and boys’ ToM. Thus, the issue of whether gender differences relate to ToM remains an open question.

The issue of gender does not necessarily trouble the thesis that language and ToM might be very closely related (Wiesmann et al., 2017). However, it does beg the wider question of whether there are alternative conceptions of the relationship between language and ToM, not just the above language facilitation view (Vierkant, 2012). For instance, Moran (2013) claims that, even if only in adults. ToM may be quite independent of cognitive domains, such as language (Wright & Wright, 2021). Other theorists (e.g., Guajardo & Cartwright, 2016) extend this thesis into childhood.
So, the apparent-place of language in children’s development of ToM may stem from it being crucial to reasoning about minds; or, language and ToM might not be related, with some third variable explaining the apparent association. However, there is a further possibility, which is that language may prohibit the child demonstrating a well-developed ToM. This follows because language tends to constitute the vessel through which ToM is predominantly assessed (Bloom & German, 2000; Milligan et al., 2007; Sax & Ranwisher, 2003; Slade & Ruffman, 2005). Even if a child’s linguistic development is a facilitator of ToM, logically that conclusion should not be regarded as safe, until the converse possibility that language might be an obstacle to ToM, has been adequately investigated (Guajardo & Cartwright, 2016; Milligan et al., 2007).

1. Aims of the research
The main goal of the present research was to select between the “language as obstacle” view and the “language as facilitator” view of children’s ToM. This was investigated via three main aims. Our first aim stems from Andrews (2005), who asserted that potential language confounds can be reduced by avoiding the need to rely heavily on language competencies (e.g., complex syntax). Responding to this critique, we indexed ToM via satisfaction (the child’s acceptance of the outcome, or surprise at the outcome), additional to only verbal predictions, in order to determine if surprise reveals higher or lower ToM.

The second aim was to explore the effect of different amounts of language on ToM, and the possibility that general language use may be sufficient to explain ToM. Here, we used children’s linguistic precociousness to determine whether it was closer to a child’s verbally expressed ToM than to our index of satisfaction of the outcomes on our ToM task.

The third aim was to determine if girls’ ToM advantage over boys’ is dependent on how much the index relies on linguistic competencies. We expected the strongest contrast between the genders to be on the ToM verbal-explanations index where the child explicitly shows their understanding of ToM, and expected the least difference on the surprise index.

Our procedure used film clips of real people acting out various object-hidings and this permitted us to ask for standard verbal responses but also to show the child the actual outcome of the films (retrievals). To allow us to consider language more closely, we included two more indexes. One was about children’s verbally explained ToM views and the other was their overall willingness to use language during the task, regardless of whether this was or was not relevant to ToM (our measure of linguistic precociousness).

Finally here, because it was important to obtain performance on the standard (verbal) task that was likely to yield a fairly symmetrical variation about the mean, something that would not occur for 3 year-olds (typically at 15–25%—Bernstein et al., 2007; Guajardo & Turley-Ames, 2004; Lockl & Schneider, 2007), this exploratory study was conducted with a 4 year-old group (Aastington, 1998).

2. Method

2.1. Participants
Participants were 55 typically developing children (30 girls) from the reception classes of three primary/infant schools in England. The schools catered for children mainly within the working class and lower middle-class socio-economic groups. The group mean age was 4.77 years (SD = 0.314; range 4.16 to 5.34 years). No children were excluded from taking part.

2.2. Materials
Eight short film clips were created depicting false-belief situations. Each clip utilized two actors, one of whom (the protagonist) placed an object in a certain location (A), and the other who attempted to deceive the protagonist by moving the object to a different place (B) whilst the protagonist was out of the room. To add variability and help keep children’s interest in the clips,
half of the clips used only female actors, with the remainder using one actor of each gender. The contexts used were a kitchen or a lounge/bedroom. In each case, three of the following potential hiding places were shown to the child. These were a kitchen cupboard, fridge or drawer; or a lounge drawer, large cupboard, or wardrobe. These were intended further to add to the variability across clips and reduce the likelihood of perseverative responses (i.e., children might otherwise tend to select the same place for where the protagonist will look). Of the three places shown to the child, only two of these at a time were used for hiding/re-hiding the object. The objects hidden were an apple, a banana, a bar of chocolate, or a can of Coca-Cola.

The various clips (approximately 1 minute long each) were mounted onto an Apple ibook G4 computer running at 1.2 GHz and having a 36 cm monitor running in full-screen mode.

These clips had one of four possible ending types. These showed the protagonist either going to the right place versus the wrong place (search), in intersection with the object being retrieved during the search versus not retrieved (outcome). Please see procedure for details on how each clip unfolded. One of each pair of endings representing an intersection of search and outcome, featured two females, with the other featuring one female and one male.

To encourage participants to rely on their interpretive skills to understand the clips, whilst leading them as little as possible (e.g., via prosody of the protagonists’ voices), each clip used no sound (images only). Earlier piloting had already shown this procedure to be suitable both for typically-developing 4 year-old children and children of 8–12 years being on the autistic spectrum.

2.3. Design

The experiment used a repeated measures design, with type of response index as the main independent variable and correct versus incorrect answers on each individual film as the main dependent variable. One index of ToM was the extent to which participants gave verbal explanations that implied they understood the nature of beliefs, a second index was the more standard verbal/pointing response and a third index was observation of whether or not the participant expressed surprise at the outcome of any trial. Separate to ToM indexes, a further index was linguistic precociousness which was used as an indication of language in regression analyses.

2.4. Procedure

The study was approved by the relevant university ethics committee prior to commencement and adhered to the British Psychological Society’s code of ethical conduct. Testing was done in a designated area of each school. Children’s ongoing assent for participation and their consent for video recording to begin, was sought. This was in addition to prior consent from a parent and from the teacher (Wright & Mahfoud, 2014). The recording was to be consulted later on, when the hand-written notes had been incomplete.

Each child was invited to watch eight film clips with a female researcher who had not featured in the clips. To use one film as an illustration of the general task: The protagonist and another actor (here both female) are in a kitchen. The protagonist has some chocolate and looks like she really wants to eat it (based on the task of Wimmer & Perner, 1983). However, she then appears to remember something and so places the chocolate in a nearby cupboard, promptly leaving the room. The other character smiles deceitfully and then goes to the cupboard, removes the chocolate and places it in a kitchen drawer instead. Watching all of this intently, the child now has a true belief about where the chocolate is. The question is, does s/he realise that the protagonist now has a false-belief about the chocolate. To find out, the clip is paused just after the protagonist has left the room.

In some false-belief studies, the experimenter asks a memory check question ahead of asking for the ToM responses. However, this may provide a shortcut cue to the correct answer Wright & Dowker, 2002; Wright & Mahfoud, 2012; and contrast Gopnik & Astington, 1988 v; Perner et al., 1987). As the
verbal-explanations response would tend to state the conflicting beliefs (see below), there was no need to additionally ask a memory question about the initial placement of the object in the video, because this initial placement would be alluded to by the child’s verbalisations. The child was now asked the ToM question as usually posed: “When xxxx comes back, where is s/he going to look for his/ her chocolate?”.

To avoid overly distracting the child, the researcher, sat facing the area between the child and the screen, and took notes from this position; which was to the side and slightly further away from the video screen than was the child. After giving his/her answer, the child was asked, “So why did you say that?” The response was taken as a verbal explanation of the ToM response. This was the most explicit index taken in the study. Finally here, we recorded participants’ verbal utterances upon each trial, as an overall measure of linguistic precociousness. This was indexed in terms of the median average number of words the child uttered to the experimenter during the eight trials of the task, irrespective of what the child was verbalising about. For instance, it might have been expression of a liking for the chocolate, or a narrative about who the child intends to play with at playtime, or even a narrative of what was happening in the silent videos (Eames et al., 1990). Linguistic precociousness was useful, because we were interested in the relationship between the child’s conversational use of language with the researcher, independent of the ToM response (Hale & Tager-Flusberg, 2003; Laranjo et al., 2010; Wang & Su, 2009). This measure was therefore mostly about each child’s willingness to talk, and the length and expressiveness of the language they used, irrespective of whether the utterance had anything to do with ToM or even the question asked. Examples of two utterances contributing to overall precociousness score during the ToM question were: “Because then they won’t take it to their home again.” versus “Because she will.”

Upon the child giving these answers (e.g., the initial answer might be given by voice or by pointing at the place), the clip was resumed and played out to the end. There were four ending types (or sub-tasks/conditions), with two film clips depicting each one. In the first ending type, when the protagonist returns s/he goes to where s/he left the object. In our example, the girl opens the cupboard to reveal no chocolate there (the facial expression is conveniently away from the child’s gaze). For the second ending, the protagonist again looks where s/he should, but the object is actually in there. This scenario violates the expected outcome. Note, “expected outcome” is the term used because it is not simply whether or not the object is present or absent; it is rather about what the expected outcome (find v not find object) should be; given that the protagonist has already gone to a specific (correct or incorrect) location to retrieve the object.

In the third and fourth ending, the protagonist goes to the place where the child expects the object is now. For ending 3, the object is indeed there; but for ending 4, it is missing. Thus, we have crossed two alternative searches (one likely and the other unlikely), with two alternative outcomes (again one likely and the other unlikely). These endings should evoke a certain amount of surprise, shown via body language and/or verbalisations about the outcome. As two examples, the child might move his/her head very close to the screen and say “Huh!”; or might turn to the researcher and say, “How did that happen?”.

During each film, the experimenter recorded and made notes on the child’s expressions, verbalisations, and behaviour, particularly in relation to amount of surprise shown when the ending was seen. Order of showing clips was randomised, and along with briefing/debriefing and familiarisation activities, the experiment took 15 to 20 minutes to administer. Children’s participation was video recorded (Scott & Roby, 2015), and each response was scored by two raters who judged whether there was surprise or no surprise. Across the 440 responses (55 participants × 8 responses each), the inter-rater reliability was 0.94. The relatively few discrepancies were resolved by discussion.
3. Results

3.1. How we scored the four composite indexes (total scores)
ToM performance was scored using three different methods. First, the standard index was taken, where children were asked to indicate where the protagonist would look for his/her object (e.g., “in the cupboard”). Because the child could simultaneously point to the location or just say as little as one word (e.g., “there” whilst pointing), this response was taken to be verbal/pointing. The standard response was either right or wrong (coded 1 v 0), depending on whether or not the child correctly indicated the place the protagonist should look.

The second index was a verbal explanation that indicated the extent to which the child was explicitly entertaining the subjective belief of the protagonist (i.e., understood ToM). Here, one mark was given for a statement that referenced minds or the subjective knowledge/belief state of the protagonist (e.g., “because that’s where she left it”). This is because the child would be intimating awareness that the protagonist should search for the object in the last place the protagonist knew it to be, in spite of the child knowing it is not there. Half a mark was awarded if the response was considered indirectly related to belief or minds (e.g., “because she put it in there”). This was because although the child might well be intending the same meaning as above, s/he might alternatively be stating a past factual observation with little direct indication of linking this to the subsequent movement of the object. The mark of 0.5 was considered a fair compromise, when we were not certain of whether the response was a statement of belief or a recollected fact.

For the third index, each child was observed for visual/oral signs of surprise upon the film clip being played to its conclusion, and the search/outcome becoming evident. A mark of 1 was given if the child showed surprise at the outcome of the clips (e.g., goes very close to the screen then says, “No way!” and finally turns to look at the researcher, or says “huh?”, or opens mouth and raises eyebrows). There was only one sub-task that was an exception to this, which was the standard ToM sub-task, when the protagonist went to the place s/he believes the object is and the object was not there. For this sub-task, additional to the above marking scheme, an additional score was calculated by reversing this measure, in order to reward the correct surprise response, which in this sub-task was to not show surprise. Thus, here, a mark of 0 was given if surprise was shown and a mark of 1 was given if no surprise was shown. The ToM scores for our three indexes were each out of 8 (corresponding to the total number of clips seen).

For children’s linguistic precociousness, we counted the number of words spoken by the child either spontaneously upon seeing the ending of a clip, or during their verbalisations in response to the questions asked to assess ToM. The median average was calculated as the child’s linguistic precociousness score. The three ToM scores and the precociousness score were deemed amenable to parametric tests. All other data/groupings were analysed with non-parametric tests suited to two or more related groups/cases. In every instance, tests were two-tailed with an alpha level of 0.05.

3.2. What we learn from analyses of means
Mean composite scores for the three ToM indexes are shown in Table 1 for girls and boys respectively. Scores are additionally converted to percentages to aid comparison with other studies. Table 1 shows a tendency for ToM scores to increase from the index where the child needed to use language the most (ToM verbal explanations), through to the index relying only on basic language labels and pointing (ToM verbal/pointing), and was highest in the index that used surprise instead of actively relying on language at all.

A two-way mixed model Analysis of Variance (ANOVA) was carried out with ToM index as the within-subject factor and gender as the between-subject factor. Note, for this analysis we used the reversed surprise scores for the two versions of film 1; because we were interested in non-surprise, which was the correct response for this surprise sub-task. In other words, this analysis considered...
the correct surprise response for each sub-task. This ANOVA revealed the overall difference between the three indexes was statistically significant, $F(2,106) = 66.545$, $p < 0.001$, Partial $\eta^2 = 0.557$, Obs.Power = 1.000. Repeated contrast analyses confirmed that the higher scores of the standard verbal/pointing index compared to the index requiring verbal explanations of ToM responses was significant, $F(1,53) = 43.109$, $p < 0.001$, Partial $\eta^2 = 0.449$, Obs.Power = 1.000. This also confirmed that scores on the surprise index were significantly higher than on the standard index, $F(1,53) = 31.804$, $p < 0.001$, Partial $\eta^2 = 0.375$, Obs.Power = 1.000.

Regarding gender, Table 1 shows that overall performance was similar for girls compared to boys. The smaller difference in favour of girls was not statistically significant, $F(1, 53) < 1$. However, there was a tendency for girls to do better than boys on verbally explained ToM but boys to do better when surprise alone was used; with no gender difference on the standard false-belief index which could be considered intermediate between the above two indexes. This two-way interaction approached statistical-significance, $F(2,106) = 2.815$, $p = 0.089$, Partial $\eta^2 = 0.045$, Obs.Power = 0.489.

### Table 1. Summary of mean ToM performance according to ToM index and gender

|                  | Verbal-explanations Index | Standard Index | Surprise Index | Overall |
|------------------|---------------------------|----------------|---------------|---------|
|                  |                           |                |               |         |
| Girls            | 2.000 (0.306) 25.0%       | 2.950 (0.385) 36.9% | 4.867 (0.267) 60.8% | 3.272 (0.216) 40.9% |
| Boys             | 1.160 (0.336) 14.5%       | 3.060 (0.421) 38.3% | 5.360 (0.292) 67.0% | 3.193 (0.237) 39.9% |
| Both             | 1.580 (0.227) 19.8%       | 3.005 (0.285) 37.6% | 5.113 (0.198) 63.9% | 3.233 (0.160) 40.5% |

Note: Maximum mean score = 8. Values in parentheses are Standard Errors. Values in italics are means expressed as percentages.

3.3. What we learned from comparisons between proportions for surprise Sub-Tasks

Having confirmed the superiority of ToM performance when assessed using surprise rather than linguistically, we turn now to analyses of the four different surprise sub-tasks. For these analyses, we used the un-reversed surprise scores, to reflect the assumption that there should be very little surprise for the standard sub-task (see C1 below). The eight film clips included two versions of each of the four surprise sub-tasks, formed by intersections of correct versus incorrect search, and correct versus incorrect outcome. Two surprise sub-tasks involved the protagonist going to the right place (the place s/he should go to): C1 = likely-search/likely-outcome (the protagonist goes to where s/he left the object and does not find it there); C2 = likely-search/unlikely-outcome (as C1 but the object is actually retrieved, which should not be what was expected). The other two sub-tasks involved the protagonist going to the wrong place (the place s/he should not know the object is now): C3 = unlikely-search/likely-outcome (the protagonist goes to where the child knows the object now is, which should have not occurred, but then retrieves it there); C4 = unlikely-search/unlikely-outcome (as C3 but the object is not there, which should have not occurred).

The standard index had been drawn from an earlier pause-point in each respective clip before the various outcomes were known, and so there should be no performance difference for the four standard sub-tasks. But for the surprise index, it was expected that a child having ToM should be more surprised when either the character searches in an unexpected place or when the object turned out not to be where the participant last saw it. In other words, there should be a difference between the four sub-tasks of the surprise index but not for sub-tasks of the standard index.
The percentages for the four standard sub-tasks are shown in the front bars of Figure 1, with the analogous sub-tasks for the surprise index shown in the back bars. We see that the four standard sub-tasks elicited very similar performance to one another, whereas the surprise sub-tasks tended to elicit rather different performance to one another. Non-parametric (Wilcoxon) tests were carried out on the four standard sub-tasks, using the Bonferroni method to adjust significance levels in view of there being six possible pair-wise comparisons. The levels of significance shown below, are the effective (equivalent) levels after applying the Bonferroni method (i.e., please divide by 6 to retrieve the p level before Bonferroni had been applied).

For the standard index, these tests indicated that there was no statistically significant difference between any single sub-task and any other single sub-task (each p > 0.05). By contrast, equivalent analyses for the analogous sub-tasks on the surprise index, revealed differences that were often statistically significant: Unlikely search/unlikely outcome (C4) differed significantly from unlikely-search/likely-outcome (C3), Z = 4.16, p < 0.05; from likely-search/unlikely-outcome (C2), Z = 3.37, p = 0.05; and from likely-search/likely-outcome (C1), Z = 4.48, p < 0.05. Note, dividing by 6 allows us to retrieve an estimate that each p ≤ 0.008 prior to applying Bonferroni correction.

By contrast, likely-search/likely-outcome (C1) did not differ from likely-search/unlikely-outcome (C2), Z = 2.77, p > 0.10; or from unlikely-search/likely-outcome (C3), Z = 1.16, p > 0.50. Finally here, likely-search/unlikely-outcome (C2) did not differ significantly from unlikely-search/likely-outcome (C3), Z = 1.62, p > 0.50.

---

Figure 1. The Surprise index always resulted in higher performance than the standard verbal/pointing index. Also, the levels for the Surprise index varied as a function of the plausibility of the location searched and retrieval outcome.

Note: C1 = Likely-Search, Likely-Outcome. C2 = Likely-Search, Unlikely-Outcome. C3 = Unlikely-Search, Likely-Outcome. C4 = Unlikely-Search, Unlikely-Outcome.
The pair-wise comparison that most closely equates to what is usually tested in developmental ToM tasks is between likely-search/likely-outcome (C1) and unlikely-search/likely-outcome (C3). This contrast is really about whether the protagonist goes to the place s/he left it but does not find it, or goes to where the other character moved it to with only the child’s knowledge, and does find it. However, the analyses above indicated that this comparison did not approach statistical significance. Four year-olds were not unduly surprised when the protagonist went to the place s/he should not have known the object had been moved to and saw it there.

Turning to the comparison between likely-search/unlikely-outcome (C2) and unlikely-search/unlikely-outcome (C4); this contrast is formally equivalent to that just presented, apart from the object not being where the events indicate it should be. Now, 4 year-olds appear very sensitive to where the protagonist searched. This, in conjunction with a review of whether children tended to use ToM-relevant language in C4, supports the interpretation that this “impossible index” was indeed revealing that children were thinking in terms of minds in C4.

We now consider an even split between the four surprise sub-tasks, according to three ways of characterising the surprise task. These are shown in Figure 2. First, the child might be interested in where the protagonist searches. Regarding correct search v incorrect search, we consider performance in terms of the surprise shown when the protagonist goes to the correct place, irrespective of whether or not the object was subsequently found therein. Please note, as the analyses here were considered alternative/independent from one another, no adjustment of significance levels was made here. The difference between likely-search/likely-outcome (C1) plus likely-search/unlikely-outcome (C2) versus unlikely-search/likely-outcome (C3) plus unlikely-search/unlikely-outcome (C4), as shown in Figure 2, was 15.5% in favour of unlikely-search. This indicates that when we in some sense controlled for object position, 4 year-olds tended to show a stronger ToM capacity. The difference between these two respective sub-task-pairs was statistically significant, Wilcoxon, Z = 2.88, p < 0.01.

Second, the child might be interested in the object itself and not so interested in the belief of the protagonist. To analyse this, we compared the presence versus absence of the object at the place the protagonist actually goes to (A or B), irrespective of whether the protagonist should actually have attempted to retrieve it from that place. We contrasted surprise for likely-search/unlikely-outcome (C2) plus unlikely-search/likely-outcome (C3) versus likely-search/likely-outcome (C1) plus unlikely-search/unlikely-outcome (C4). From Figure 2 we see that unlikely-outcome led to more surprise than likely-outcome; and the difference was far greater than for the previous analysis. Figure 2 shows that children were far more interested in whether the object was found than

---

Figure 2. The biggest difference between the bar for the likely situation compared to the corresponding bar for the unlikely situation did not occur when participants were reasoning about others’ minds (i.e., search), but occurred instead when considering where the object should have been really (i.e., outcome).

Note: Search = Likely v unlikely irrespective of outcome; Outcome = Likely v unlikely irrespective of search; Exposure = Exposed v did not expose object at location search.
whether the object was where the protagonist thinks it should be. The 27.25% difference as shown in Figure 2 was statistically significant, Wilcoxon, Z = 4.57, p < 0.01.

The third analysis here was about whether the child was interested in the outcome of searches, regardless of whether the protagonist should believe the object should be in any particular place (A or B). Considering the correct versus the incorrect outcome, should help us determine whether our directly preceding finding indeed reflects an interest in the object, or may simply reflect that the child has an interest in whether the search produces the expected outcome (from the protagonist’s point of view). We contrasted the sub-task-pair likely-search/likely-outcome (C1) plus unlikely-search/likely-outcome (C3) versus likely-search/unlikely-outcome (C2) plus unlikely-search/unlikely-outcome (C4). From Figure 2 we see that this contrast yielded the smallest difference. Indeed, the 7.25% difference was not statistically significant, Wilcoxon, Z = 1.47, p > 0.10.

3.4. What we learned from correlational and regression analyses (at the individual level)
Our next analyses first considered the relationship between the three overall ToM indexes. A set of Pearson’s correlations was calculated to consider the pairwise associations between the standard verbal/pointing ToM index, verbal-explanations index, and surprise index, respectively. Children’s linguistic precociousness was added as a further variable, because of previously reported findings of links between language or conversational abilities and ToM. Gender was added as the final variable here because of the tendency towards a significant interaction between gender and the ToM index being considered (particularly the verbal explanation of ToM responses).

Pairwise correlations are summarised in Table 2. Table 2 showed that the surprise index did not correlate well with any other variable, including gender. The standard verbal/pointing index of ToM did not correlate directly with linguistic precociousness. However, both these variables did correlate with the verbal-explanations ToM index, as did gender. The sign of the correlation between gender and the ToM verbal-explanations index was negative, indicating that girls (coded as 1) tended to score more highly than did boys (coded as 2).

A regression analysis was now conducted to analyse the extent to which the surprise index and the more standard verbal/pointing index, predicted ToM performance based on verbal-explanations. The verbally explained ToM responses were chosen as the criterion, because it had been by far the most challenging ToM index to the children. As we needed to consider whether a child’s linguistic precociousness could account for verbally explained ToM, we added the measure of linguistic precociousness as a further predictor. Recall this measure was about children’s general talkativeness, irrespective of whether they alluded to ToM or even addressed the question they were asked, when they replied following each question. Next, the very first analysis (mixed-model ANOVA) had suggested that there was a tendency for performance to favour girls the more the ToM index involves language; which approached but did not reach statistical-significance. However, the correlational (individual-based) analyses showed gender was significantly associated with verbally explained ToM on a child-by-child basis. Therefore, gender was entered as the final predictor here.

As we were cognizant of our sample size being slightly moderate compared to the number of predictors, we opted for the forward stepwise method. This analysis essentially settles on the fewest variables that lead to a reliable model, which therefore increases the ratio of N-participants to N-predictor-variables. The regression is summarised in Table 3. The model R was 0.738 and this model was statistically significant, F(4,54) = 20.381, p < 0.001. The model accounted for 51.8% of the variability in scores for verbally explained ToM (R² = 0.518).

From Table 3 it can be seen that both linguistic precociousness and the standard verbal/pointing index of ToM were statistically significant predictors of verbally explained ToM, with the standard ToM index having more than twice the predictive power of the linguistic precociousness variable.
|                | ToM-Explanation | Lang Precociousness | ToM Standard | ToM Surprise | Gender |
|----------------|-----------------|---------------------|--------------|--------------|--------|
| ToM-Explanation| -               | 0.308 (0.011)       | 0.636 (<0.001) | -0.118 (0.195) | -0.246 (0.035) |
| Lang Precociousness | -           | -                   | 0.028 (0.419)   | -0.076 (0.292)   | -0.086 (0.266)   |
| ToM Standard     | -               | -                   | -             | -0.163 (0.117)   | 0.026 (0.424)    |
| ToM Surprise     | -               | -                   | -             | -             | 0.169 (0.109)    |
| Gender           | -               | -                   | -             | -             |        |

Note: ToM-explanation = Verbally explained ToM, Lang Precociousness = Linguistic precociousness, ToM standard = Standard verbal/pointing ToM index, ToM surprise = Surprise index of ToM. The values in parentheses are levels of significance for each correlation.
The final analyses were carried out to consider the reasons why the surprise index, although revealing the highest ToM ability, had not been a reliable predictor of children’s understanding of ToM as shown in their responses and verbal-explanations. We ran a preliminary reliability analysis for the four conditions of the surprise task (C1-C4) in order to determine whether they worked together to index the same construct (i.e., total surprise score). This analysis resulted in a moderate reliability estimate (Cronbach’s Alpha) of 0.59. Additionally, the removal of any one of the four conditions either led to no change in the Cronbach’s Alpha value or led to its reduction; indicating the four conditions fit reasonably into a single construct.

A follow-up regression was then carried out to determine the extent to which each of the surprise conditions predicted verbally justified ToM score, in context of the other surprise subtasks. As before, gender was included in this model and we opted for the stepwise method. The regression is summarised in Table 4.

The model R was 0.419 and this model was statistically significant, F(2,54) = 5.546, p = 0.007. This model accounted for 17.6% of the variability in scores for verbally explained ToM (R² = 0.176). From Table 4 it is evident that gender was a statistically significant predictor as before. However, the only significant predictor from the four surprise sub-tasks was C3 (unlikely-search/likely-outcome).
4. Discussion

To determine if 4 year-olds revealed different levels of ToM when we do versus do not test via language, we used stimulus presentations based on real-life situations, and depicted ToM situations using video clips of real persons to further aid realism. Under these conditions, children performed below chance on the standard linguistic index but substantially above chance on a surprise index tied to exactly the same task as the linguistic index. However, comparisons of surprise between the four conjunctions of where the protagonist searches for the object versus the outcomes of those protagonist searches, contextualised and enriched this finding. These alternative ways of estimating 4 year-olds’ ToM are discussed in turn below.

4.1. Comparing scores on our three ToM indexes

Studies of implicit versus explicit ToM reasoning tend to have been carried out in different papers or in different experiments within the same paper (e.g., San Juan & Astington, 2017). However, relatively few studies have employed within-participant designs within a single experiment (Nijhof et al., 2016; Thoermer et al., 2012). In the latter tradition, the present study incorporated at least three within-subject variables (Rosenblau et al., 2015).

Our standard (verbal/pointing) index yielded an estimate of around 37.6% for 4 year-olds. This is quite similar to estimates from several other studies (Guajardo & Turley-Ames, 2004; Lockl & Schneider, 2007; Lohmann et al., 2005; Wright & Mahfoud, 2012). When assessed via a composite ToM score based around largely bypassing language (here using surprise), children’s false-belief scores approached twice the magnitude compared to our own more standard verbal index for assessing ToM (Andrews, 2005; Yott & Poulin-Dubois, 2016). This finding is in line with Scott and Roby (2015), who showed that 3 year-olds and early 4 year-olds perform better in a task using preferential looking, as compared to when the false belief task relies quite directly on language.

Here, the advantage of the surprise index was found to be much higher when we replaced the standard language/pointing index with the need for the child to give linguistic (verbal) explanations for their ToM responses. These verbal explanations needed to infer desires, mental states, or beliefs (e.g., “Because that is where she left it”; “That's where she thinks it is but it has been moved”). Now, with this verbal-explanation index, which most explicitly captured the fact that the child’s response had been given because of the child considering the false-belief of the protagonist, the ToM estimate was a lower 19.8%. This is less than one-third of the value found for the surprise index.

4.2. Language and gender effects

From their finding that language training improved performance both on later language tasks and on later ToM tasks, Hale and Tager-Flusberg (2003) had concluded that this is evidence that language (use of sentential complements) is important in the development of ToM. The present findings cannot refute this possibility outright (Durrieu-Man-Tame et al., 2017; De Mulder et al., 2019). However, we would argue that an explanation of our own findings in terms of language indexes suggesting ToM is less developed at 4 years than it really is, also deserves additional empirical attention.

So, the present data are parsimonious with the view that language and ToM are relatively independent entities, which do not necessarily have to be reliant on each other (Bloom & German, 2000; Moran, 2013). This interpretation can actually be consistent with Hale and Tager-Flusberg’s (2003) additional finding that ToM training did not improve children’s subsequent performance on language tasks; whereas language training on sentence complements improved both later language and later ToM performance. In the present study, the lower performance on verbally explained ToM compared to the more standard index where one only has to indicate the verbal label of the place where the protagonist should go to look for his/her object; would seem to
suggest that the more the index calls on (underdeveloped) linguistic competencies (e.g., grammatical or syntactic development), the lower young children’s ToM competence would appear to be as interpreted through that index (Baillargeon et al., 2010; Bloom & German, 2000; De Mulder et al., 2019).

The finding of no reliable overall gender difference across the three indexes (Laranjo et al., 2010; Longobardi et al., 2017; Melinder et al., 2006), contributes to our confidence that our ToM measures were appropriate. However, we found that girls tended to do better than boys the more language featured in the index (Walker, 2005; Wright & Mahfoud, 2014), but not when the role of language was minimised (i.e., in our surprise index). This contrast in ToM profiles according to language, further adds to our belief that each of our three ToM measures was valid in its own right: It is in line with previous findings that male children and adolescents tend to lag slightly behind their female counterparts both in linguistically assessed ToM and in aspects of language itself, such as linguistic precociousness, reading, and spelling performance, use of conversational language or even talking about emotions (Hale & Tager-Flusberg, 2003; Laranjo et al., 2010; Wang & Su, 2009; Wright & Mahfoud, 2014; Wright & Wright, 2021). If, as we believe, language does moderate the extent to which a child’s ToM can be seen, rather than (or as well as) actually causing ToM, is correct, one would expect this finding regarding differences to emerge the more the tasks we use draw on language. In their recent study of ToM, Wright and Wright (2021) confirmed this contrast in regard to one disability thought to have a linguistic component—dyslexia. They found that an adult variant on the false-belief task showed lower performance for young adults having dyslexia; however, when this was replaced by a task far less reliant on linguistic processes, differences between the group having versus not having dyslexia were no longer evident. Wright and Wright interpreted this contrast in findings as evidence that the linguistic demands of a ToM task can indeed lead to an unfair and potentially devastating (invalid) conclusion about a particular disabled group.

When analyses were based on individuals rather than averaged across groups and indexes, we further found that gender is one of the predictors of ToM understanding when this is indexed via verbally explained subjective beliefs (Charman et al., 2002). This finding raises the possibility that one reason why there may be a slight ToM advantage for girls, particularly at younger ages (Walker, 2005; Wright & Mahfoud, 2014), is because the standard task of false-belief that is often taken as the measure of children’s ToM competence, tends to have been assessed via language. On this account, perhaps investigators should take more seriously the possibility that it may be language rather than a genuine ToM competence that yields a slight advantage to girls. Consistent with this view, whereas the two tasks more heavily relying on language in the present study did indicate an advantage to the ToM of girls, the surprise task showed no such advantage.

4.3. The nature of ToM as indexed by surprise

Focussing on the three main ToM indexes used here (verbally explained beliefs, linguistically reporting the place the protagonist will look, and amount of surprise upon seeing the conclusion of the protagonist’s search for the object), seems to suggest that it may be that young children reveal different levels of ToM, depending on the response mode utilized (Leslie et al., 2005). Interestingly, the linguistic precociousness variable was found to be less related to verbally explained ToM, than was the strength of relationship between the standard verbal/pointing index and verbally explained ToM. This was the case both in the correlational analysis and in the regression analysis of total scores. Thus, it seems that verbally explained ToM was caused more by a genuine ability to understand minds than by linguistic precociousness.

If we now consider the first two ToM indexes to reflect varying degrees of language competencies and contrast this against the surprise index, this too suggests there are at least two modes of ToM, with one being verbal/explicit and the other less-verbal or even non-verbal/implicit. The implicit mode is said to rely on the same neural structures as the explicit mode (Naughtin et al.,
2017), but developing earlier (Clements & Perner, 2001; Etel & Slaughter, 2019; Low, 2010; De Mulder et al., 2019; Yott & Poulin-Dubois, 2016).

This implicit-explicit distinction may explain conflicting age estimates for ToM-like abilities in infancy (Kloo et al., 2020). To illustrate, consider again Onishi and Baillargeon’s (2005) finding of ToM-like abilities by age 15 months, as contrasted with Repacholi and Gopnik’s (1997) null result for infants of similar age. The apparent conflict may simply reflect when children’s responses to ToM questions are based around linguistic utterances from the experimenter (e.g., comprehending by 19 months—“I’d like some more please”). Whereas if telling the story non-linguistically and taking looking times instead of linguistic predictions/explanations as the index of ToM, the estimates should now be earlier (hence 15 months in Onishi & Baillargeon, 2005).

The present correlational and the initial regression analyses which concerned the total scores on each index, seemed to support the view that the surprise index, although it indicated a much higher level of ToM than either of our two language-laden indexes, was itself not reliably related to the most explicit language indexes (Rosenblau et al., 2015; Wiesmann et al., 2017). However, a key feature of implicit ToM is that the experimenter infers ToM data by indirect means (Low & Perner, 2012). This means the reasoner need not be aware that s/he is giving response profiles that imply ToM competencies (e.g., looking times, time for anticipatory looking, first look, preference or RT indexes—Baillargeon et al., 2018; Kloo et al., 2020; Naughtin et al., 2017; Nijhof et al., 2016; Scott & Roby, 2015; Thoermer et al., 2012).

When people do show surprise that surprise tends to occur relatively fast and the reasoner may not even be consciously aware of the response until after it had been given (Kloo et al., 2020). So should we consider the surprise index to be implicit, simply because it is not explicit? We argue the correct answer is “no”. This is because, although not explicit, the surprise responses do not represent an indirect index that would be required to apply the label “implicit” (Low & Perner, 2012).

We may draw some support for this view, by considering the individual-based analyses. The first regression showed that the surprise index did not predict scores on verbally explained ToM. Reliability analyses did confirm that the four surprise conditions showed adequate internal reliability, and hence can be regarded as indexing a single construct (of surprise). Therefore, it would seem that surprise, as a unitary index, was distinct from the language index. We then carried out a second regression analysis to determine if any of the surprise sub-tasks do predict the language index of ToM. This indicated that only the sub-task Unlikely-Search/Likely-Outcome (C3) reliably predicted the total score on the most demanding ToM index. This is where the protagonist goes to place B to search for the object, behaviour, which could not have been based on the protagonist’s belief (which should have led to a search at place A).

The more the child showed surprise when the protagonist was going to retrieve the object from the location the protagonist did not see that object being placed, the better the child did on the most explicit ToM index that included the child explaining why the protagonist should have gone to the place the protagonist had last seen the object. This finding makes complete sense if we regard our surprise measure as a second explicit index of ToM. However, if our surprise measure of ToM is taken to be an implicit index, it becomes more difficult to explain.

Indeed, when adults show surprise, we do not tend to interpret that response as implicit. This is especially true if the surprise is accompanied by spontaneous verbalisations and/or conscious exclamation (Vierkant, 2012). For instance, the reasoner might say “Huh!”; “how did that happen?”; followed by leaning in really close to the computer screen or looking wide-eyed at the experimenter. The point is, these behaviours tend to be at least as explicit as they are implicit. The upshot is that our surprise index is an intermediary between implicit and explicit ToM responding (Thoermer et al., 2012).
As implied by the above discussion on C3 in the individual-based (regression) analysis, even within our surprise index itself, we again obtained different levels of ToM (see, Fabricius & Imbens-Bailey, 2000) for anticipation of such a finding. For example, in C4, where the protagonist is seen going to where s/he did not leave the object but then does not retrieve the object; this is actually the right outcome because the protagonist should not have found the object where s/he should have searched (which should have been at A). But at the same time it is an unlikely search location in that s/he should in any event not have gone to that particular place to search for the object anyway. Intriguingly, despite surprise in C3 being the sub-condition that most predicted ToM as indexed linguistically, it was C4 that led to greatest surprise; and yet linguistic precociousness for the complete sub-condition was identical in C3 and C4. Note, both in C3 and C4, the protagonist retrieves the object at B, which is where the child but not the protagonist should know the object was last put. However, at this time, it is difficult to disentangle an account of the contrasts between surprise in C3 and C4 in terms of the child’s own knowledge being violated versus the protagonist’s search being unsuccessful; versus ceiling effects on C4 but not on C3. Future research will be required on this issue.

Our group-based analyses using the splits for sub-task-pairs indicated two related findings. First, our child participants showed appropriate surprise when the protagonist went to search at the wrong place (irrespective of whether the object will be retrieved there). However, they were far more attuned to whether the object was retrieved or not (the outcome), than whether the protagonist went to where s/he should have gone or not (belief/intention). Repacholi and Gopnik (1997) obtained an analogous finding with their older infants. In their experiment with food-preferences, both their 19 month-olds and their 14 month-olds gave the experimenter the correct food when the child him/herself had also preferred that food, more often than when the child had not liked that food. This finding and our own here, is consistent with findings for non-human primates, tested in a situation where they can choose a reward just for themselves or both for themselves and another participant (Silk et al., 2005). They are also consistent with what has been termed a “realist bias” (Flavell et al., 1985; Leslie et al., 2005). The realist bias basically refers to the reasoner giving primacy to information held as real/current, over information that either is outdated knowledge held by someone else or previously held by oneself (see also hindsight bias—Bernstein et al., 2007; Birch & Bernstein, 2007).

Our second finding here was that children found it much easier to respond according to the knowledge/intention of the protagonist in the film clips, when the subsequent outcome of the search was impossible. Indeed, when we restricted ourselves only to the possible (plausible and expected) locations of the objects, children were no more surprised by the protagonist going to where the object is now, compared to him/her going to where s/he had actually left it. In other words, we now observed rather poor ToM at 4 years. This finding clearly demonstrates that even children failing to show any expectation about where a protagonist will search for an object whose location has been changed, might nevertheless have the ability to reason about minds under the right conditions (Repacholi & Gopnik, 1997). The fact that here the appropriate condition was when the expected outcome regarding the object was violated, qualifies our previous conclusions on a realist bias, and also confirms that children may have the right mental mechanisms for ToM well before they reliably use this mental mechanism to reason about real minds in real situations (Bloom & German, 2000; Kloo et al., 2020; Leslie et al., 2005; Low, 2010; Onishi & Baillargeon, 2005).

In the limit, it may even be that ToM is distinct from the cognitive/linguistic abilities that are associated with the reporting of false-belief but tends to appear absent in the under-5s because false-belief tasks typically access it via language, which may be less developed than the young child’s ToM competencies. This conclusion regarding language versus ToM becomes clearer if we consider that, in everyday life, for the most part neither children nor adults need routinely to explain or even merely verbalise their ToM reasoning (Ruffman, 2014).
So, as Hobson (1994) intimates, we should be investigating the possibility that ToM may even be innate. It may need only permission from a more socially experienced other, in order to move closer and closer to adult levels (Hobson, 1994). Thus, what may actually need to develop is, not so much ToM itself, but rather the child’s realisation that older children and adults want him/her to engage this mechanism linguistically and apply it to real situations. Hobson’s point notwithstanding, it may yet be that the most intuitive way for testing explicit ToM in children is still the use of a false-belief task. But perhaps tasks should judge children’s reactions to situations with varying levels of implausibility, alongside asking for their verbal responses or verbal-explanations.

4.4. Potential limitations of the study
As with all studies, the findings of this research should be viewed in context of some potential limitations. The main one may be that the sample size of fifty-five 4 year-olds could be considered rather modest. However, if sample size was a limitation, this would render the positive findings discussed above all the more compelling. Also, this number of children is quite similar or even larger than samples or experimental groups in some classical and recent studies (Baron-Cohen et al., 1985; Kloé et al., 2020; Scott & Roby, 2015; Thoermel et al., 2012; Wiesmann et al., 2017). This point also holds in comparisons against some studies of ToM in adulthood (Nijhof et al., 2016; Rosenblau et al., 2015). To minimise sample size being an issue, the present research used a false-belief task where each child gave a total of 8 responses for each of the three main indexes (surprise, standard verbal/pointing, ToM verbal-explanations), totalling up to 24 times the data from each child, compared to other studies such as some of those cited above.

Secondly, it could be argued that in the surprise task, only the condition where the child saw the protagonist going to the place s/he left the object (A) but the outcome was a failure to retrieve the object (C1), and the condition where the child sees the protagonist go to the place the protagonist had not left the object but now the outcome is that the protagonist retrieves the object by searching at B (C3), are relevant to ToM. On this issue, we would prefer to await additional empirical evidence. But the fact that the four surprise sub-conditions reliably combined into a single construct, does support our use of all four of these conditions in the data and analyses.

A third potential limitation is our reliance on only one measure of linguistic precociousness. This was mainly to make most effective use of the valuable time that the schools permitted us to work with their children during school time. Although we regard the measure we did use both as relevant and informative, we accept that we could have added observations of children’s language use in the classroom and the playground. Alternatively, we could have employed separate vocabulary, syntactic, comprehension, or oral tasks as part of our experimental procedure (Low, 2010; Miller, 2006; De Mulder et al., 2019; Slade & Ruffman, 2005). Including such tasks alongside the spontaneous index we used, would permit us to cross-validate our linguistic precocity index against other indexes of linguistic use.

We note that our own linguistic precociousness index was a reliable predictor of verbally explained ToM and hence was an adequate index. This made it all the more intriguing that linguistic precociousness was not correlated with the standard ToM verbal/pointing index, which itself tended to include a verbal component. This profile potentially has many possible interpretations. Thus, further research will need to consider this issue. However, we are satisfied well enough with the use of linguistic precociousness here and we would recommend such a measure be included in future studies investigating relationships or interactions between language and ToM.

Finally, although earlier we consider our interpretation of the implicit versus explicit dichotomy in terms of a gradation of ToM-like competencies rather than two independent phenomena, it is accepted here that there may well be other alternative explanations of our findings regarding the three indexes of ToM we used (Guajardo & Cartwright, 2016). But even in that event, the present findings will have stimulated a deeper understanding of ToM development. It is our view that relying on continuity or intermediary modes of exhibiting ToM responses, over-and-above the
dichotomous implicit-explicit distinction (Couchman et al., 2012), could help elucidate the competencies of groups considered atypical. For example, certain groups such as deaf or blind children, when tested using tasks originally tuned to response modes of sighted children, may tend not to reveal their true ToM competencies (Lecciso et al., 2016; Roch-Levecq, 2006). This potentially can explain why existing research findings suggest that acquiring ToM for these groups can take up to three times as long as for neuro-typical children (i.e., ToM arrives between 10 and 12 years in blind children—Peterson et al., 2000). What is required now are studies re-evaluating this thesis in relation to blind children and other key groups.

5. Conclusions
The standard false-belief task, which typically employs linguistic-verbal responses, does deal well with the possibility that language might play an important role in ToM. Two such indexes here, confirmed that ToM is not all that well developed by 4 years. We also confirmed that the greater is the reliance of an index on linguistic abilities, the more likely that 4 year-old girls will outperform boys. This is in line with previous research that has concluded language is an essential driver of ToM development.

However, we additionally considered the possibility that language may limit the extent to which younger children can show us their well-developed ToM competencies, even if we accept the proposition that language in some respects also facilitates ToM. When we used a composite index relatively free of language (our surprise index), we found that ToM is far more developed than shown using the standard language/pointing index or shown via verbal explanations in terms of beliefs. Indeed, performance on our surprise index was two to three times as high compared to one or other of our more language-laden indexes. This surprise index, therefore, may be particularly useful with children even younger than 4 years and children having atypical developmental profiles.

In more fine-grained analyses, we compared the amount of surprise as a function of different intersections between searches and outcomes with our surprise index. These additionally revealed that 4 year-olds demonstrate different levels of ToM appreciation depending on the interaction between expected and unexpected search strategies versus expected and unexpected outcome (finding v not finding the object). Most interestingly, when we used predictive analyses we found that the surprise sub-task where the protagonist goes to the wrong place but rightly picks the object up from there, was the only sub-task that predicted variability in ToM involving explanations, even though it was not the sub-task eliciting highest surprise.

By contrast, in the comparisons of proportions, we now found that children were better at ToM reasoning when the situation was maximally implausible (i.e., unlikely search plus unlikely outcome), compared to the much greater plausibility of the standard false-belief task (likely search plus likely outcome). They were also more interested in the presence or absence of the object, than the search behaviour of the protagonists, demonstrating more interest in outcomes than in minds. Thus, 4 year-olds do have a ToM competence but it is not yet a unitary one. What they need to do is learn about the situations to which it can advantageously be engaged, for example, by experiencing other people’s actual actions in the real world.

Acknowledgements
Sincerest thanks to Alison Ledger for assistance with creating materials and collecting data, and Karen Chapman and Amanda Wright for assistance with figures and tables. But the most special thanks are reserved for all the children who patiently took part in the study, their parents and the schools where we tested.

Funding
The author received no direct funding for this research.

Author details
Barlow C Wright 1

E-mail: barlow.wright@ntu.ac.uk
ORCID ID: http://orcid.org/0000-0002-4946-9761

1 Department of Psychology, Nottingham Trent University, Nottingham, UK.

Disclosure statement
No potential conflict of interest was reported by the author(s).

Data availability statement
The dataset on which this article is based can be obtained from the first author upon reasonable request. barlow.wright@ntu.ac.uk
Citation information
Cite this article as: Language can obscure as well as facilitate apparent-Theory of mind performance: part 1 - An exploratory study with 4 year-olds using the element of surprise, Barlow C Wright, Cogent Psychology (2022): 9: 2111838.

References
Andrews, K. (2005). Chimpanzee theory of mind: Looking in all the wrong places? Mind & Language, 20(5), 521–536. https://doi.org/10.1111/j.1082-7096.2005.00298.x
Astonishing, J. W. (1998). Theory of mind, Humpty Dumpty, and the icebox. Human Development, 41(1), 30–39. https://doi.org/10.1159/000022566
Baillargeon, R., Buttelmann, D., & Southgate, V. (2018). Invited commentary: Interpreting failed replications of early false-belief findings: Methodological and theoretical considerations. Cognitive Development, 45-48, 112–124. https://doi.org/10.1016/j.cogdev.2018.06.001
Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. Trends in Cognitive Science, 14(3), 110–118. https://doi.org/10.1016/j.tics.2009.12.006
Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a “theory of mind”? Cognition, 21(1), 37–46. https://doi.org/10.1016/0010-0277(85)90022-8
Beaudoin, C., Leblanc, E., Gagnon, C., & Beauchamp, M. H. (2020). Systematic review and inventory of theory of mind measures for young children. Frontiers in Psychology, 10(2905), 1–23. https://doi.org/10.3389/fpsyg.2019.02905
Bernstein, D. M., Atance, C., Melzoff, A. N., & Loftus, G. R. (2007). Hindsight bias and developing theories of mind. Child Development, 78(4), 1374–1384. https://doi.org/10.1111/j.1467-8624.2007.01071.x
Birch, S. A. J., & Bernstein, D. M. (2007). What can children tell us about hindsight bias: A fundamental constraint on perspective-taking? Social Cognition, 25(1), 98–113. https://doi.org/10.1521/soco.2007.25.1.98
Blooom, P., & German, T. (2000). Two reasons to abandon the false belief task as a test of theory of mind. Cognition, 77(1), 825–831. https://doi.org/10.1016/s0010-0277(00)00096-2
Coputi, M., & Schoenborn, H. (2018). Theory of mind and internalizing symptoms during middle childhood and early adolescence: The mediating role of coping strategies. Cogent Psychology, 5(1), 1487270. https://doi.org/10.1080/23311908.2018.1487270
Charman, T., Ruffman, T., & Clements, W. (2002). Is there a gender difference in false belief development? Social Development, 11(1), 1–10. https://doi.org/10.1111/1467-9507.00183
Clements, W. A., & Perney, J. (2001). When actions really do speak louder than words but only implicitly: Young children’s understanding of false belief in action. British Journal of Developmental Psychology, 19(3), 413–432. https://doi.org/10.1348/026151001166182
Couchman, J. J., Beran, M. J., Coutinho, M. V. C., Boomer, J., Zakrzewski, A., Church, B., & Smith, J. D. (2012). Do actions speak louder than words? A comparative perspective on implicit versus explicit meta-cognition and theory of mind. British Journal of Developmental Psychology, 30(1), 210–221. https://doi.org/10.1111/j.1467-8624.2011.00265.x
De Mulder, H. N. M., Wijnen, F., & Coopmans, P. H. A. (2019). Interrelationships between Theory of Mind and language development: A longitudinal study of Dutch-speaking kindergartners. Cognitive Development, 45-48, 67–82. https://doi.org/10.1016/j.cogdev.2019.03.006
Durrieu-Man-Tame, S., Burnel, M., & Rebull, A. (2017). Connections among complementation sentences, executive functioning, and theory of mind in autism. In L. R. Naigles (Ed.), Language and the human life-span series. Innovative investigations of language in autism spectrum disorder (pp. 163–182). American Psychological Association; Walter de Gruyter GmbH. https://doi.org/10.1037/15964-009
Eaves, D., Shorrock, D., & Tomlinson, P. (1990). Naughty animals or naughty experimenters? Conservation accidents revisited with video stimulated commentary. British Journal of Developmental Psychology, 8(1), 25–37. https://doi.org/10.1111/j.1467-9211.2004.835X.1990.tb00819.x
Etel, E., & Sloughter, V. (2019). Theory of mind and peer cooperation in two- and three-year-olds contexts. Journal of Applied Developmental Psychology, 60-65, 87–95. https://doi.org/10.1016/j.appdev.2018.11.004
Fabricius, W. V., & Imbens-Bailey, A. (2000). False beliefs about false beliefs. In P. Mitchell & K. J. Riggs (Eds.), Children’s reasoning and the mind (pp. 267–280). Psychology Press.
Flavell, J. H., Flavell, & Green, F. L. (1985). The effects of question clarification and memory aids on young children’s performance on appearance-reality tasks. Cognitive Development, 21, 127–144. https://doi.org/10.1016/S0885-2014(87)90104-3
Gopnik, A., & Astington, J. W. (1988). Children’s understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. Child Development, 59(1), 26–37. https://doi.org/10.2307/1130386
Guajardo, N. R., & Cartwright, K. B. (2016). The contribution of theory of mind, counterfactual reasoning, and executive function to pre-readers’ language comprehension and later reading awareness and comprehension in elementary school. Journal of Experimental Child Psychology, 144, 27–45. https://doi.org/10.1016/j.jecp.2015.11.004
Guajardo, N. R., & Turley-Ames, K. J. (2004). Preschoolers’ generation of different types of counterfactual statements and theory of mind understanding. Cognitive Development, 19(1), 53–80. https://doi.org/10.1016/j.cogdev.2003.09.002
Guise, K., Kelly, K., Romanowski, J., Vogeley, K., Platek, S. M., Muny, E., & Cutting, L. (2007). The anatomical and evolutionary relationship between self-awareness and theory of mind. Human Nature, 18(2), 132–142. https://doi.org/10.1007/s11090-009-9009-x
Hale, C. M., & Tager-Flusberg, H. (2003). The influence of language on theory of mind: A training study. Developmental Science, 6(5), 346–359. https://doi.org/10.1111/1467-7687.00289
Hobson, R. (1994). On developing a mind. British Journal of Psychiatry, 165(5), 577–581. https://doi.org/10.1192/bjp.165.5.577
Hughes, C., Deater-Deckard, K., & Cutting, A. L. (1999). “Speak roughly to your little boy”: Sex differences in the relations between parenting and preschoolers’ understanding of mind. Social Development, 8(2), 143–160. https://doi.org/10.1111/1467-9507.00088
Kloo, D., Kristen-Antonow, S., & Sodian, B. (2020). Progressing rom an implicit to an explicit false belief understanding: A matter of executive control? International Journal of Behavioral Development, 44(2), 107–115. https://doi.org/10.1177/0165025419850901
Knutsen, J., Mandell, D. S., & Frye, D. (2017). Children with autism are impaired in the understanding of
false-belief understanding. Cognitive Development, 41, 19–32. https://doi.org/10.1016/j.cogdev.2016.12.003

Sax, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in “theory of mind”. Neuronimage, 19(4), 1835–1842. https://doi.org/10.1016/S1053-8119(03)00230-1

Scott, R. M., & Roby, E. (2015). Processing demands impact 3-year-olds’ performance in a spontaneous-response task: New evidence for the processing-load account of early false-belief understanding. PiJOS One, 10(11), 1–20. https://doi.org/10.1371/journal.pone.0142405

Silk, J. B., Brosnan, S. F., Vonk, J., Henrich, J., Povinelli, D. J., Richardson, A. S., Lambeth, S. P., Mascaro, J., & Schapiro, S. J. (2009). Chimpanzees are indifferent to the welfare of unrelated group members. Nature, 437 (7063), 1357–1359. https://doi.org/10.1038/nature04243

Simmons, F., & Singleton, C. (2000). The reading comprehension abilities of dyslexic students in higher education. Dyslexia, 6(3), 178–192. https://doi.org/10.1002/1099-9069(200007)6:3

Slade, L., & Ruffman, T. (2005). How language does (and does not) relate to theory of mind: A longitudinal study of syntax, semantics, working memory and false belief. British Journal of Developmental Psychology, 23(1), 117–141. https://doi.org/10.1348/026151004X21332

Southgate, V., Senju, A., & Csibra, G. (2007). Action anticipation through attribution of false belief by 2-year-olds. Psychological Science, 18(7), 587–592. https://doi.org/10.1111/j.1467-9280.2007.01944.x

Talwar, V., Crossman, A., & Wyman, J. (2017). The role of executive functioning and theory of mind in children’s lies for another and for themselves. Early Childhood Research Quarterly, 41 (4), 126–135. https://doi.org/10.1016/j.ecresq.2017.07.003

Tankersley, D., Stowe, C. J., & Huettel, S. A. (2007). Altruism is associated with an increased neural response to agency. Nature Neuroscience, 10(2), 150–151. https://doi.org/10.1038/nn1833

Theormer, C., Sodian, B., Vuori, M., Perst, H., & Kristen, S. (2012). Continuity from an implicit to an explicit understanding of false belief from infancy to preschool age. British Journal of Developmental Psychology, 30(1), 172–187. https://doi.org/10.1093/bjdp/ngr004

Vierkant, T. (2012). Self-knowledge and knowing other minds: The implicit/explicit distinction as a tool in understanding theory of mind. British Journal of Developmental Psychology, 30(1), 141–155. https://doi.org/10.1093/bjdp/ngr004

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

Wang, Y., & Su, Y. (2009). False belief understanding: Children catch it from classmates of different ages. International Journal of Behavioral Development, 33 (4), 331–336. https://doi.org/10.1177/0165025409104525

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

White, S., Hill, E., Hoppé, F., & Frith, U. (2009). Revisiting the strange stories: Revealing mentalizing impairments in autism. Child Development, 80(4), 1097–1117. https://doi.org/10.1111/j.1467-8624.2009.01319.x

Wiesmann, C. G., Friederici, A. D., Singer, T., & Steinbeis, N. (2017). Implicit and explicit false belief development in preschool children. Developmental Science, 20(5), e12445. https://doi.org/10.1111/desc.12445

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

Wright, B. C., & Dowker, A. D. (2002). The Role of Cues to Differential Absolute Size in Children’s Transitive Inferences. Journal of Experimental Child Psychology, 81(3), 249–275. https://doi.org/10.1006/jecp.2001.2653

Wright, B. C., & Mahfoud, J. (2012). A child-centred exploration of the relevance of family and friends to theory of mind development. Scandinavian Journal of Psychology, 53(1), 32–40. https://doi.org/10.1111/j.1467-9450.2011.00920.x

Wright, B. C., & Mahfoud, J. (2014). A teacher-centered exploration of the relevance of social factors to theory of mind development. Scandinavian Journal of Psychology, 55(1), 17–25. https://doi.org/10.1111/sjop.12085

Wright, B. C., & Wright, B. A. L. (2021). Language can obscure as well as facilitate apparent Theory of Mind performance: Part 2 – The case of dyslexia in adulthood. Frontiers in Psychology: Cognitive Science, 12, 621457. https://doi.org/10.3389/fpsyg.2021.621457

Yott, J., & Poulin-Dubois, D. (2016). Are infants’ theory of mind abilities well integrated? Implicit understanding of intentions, desires, and beliefs. Journal of Cognition and Development, 17(5), 683–698. https://doi.org/10.1080/15248372.2015.1086771
