Two-Year-Olds' Eye Movements Reflect Confidence in Their Understanding of Words

Isabelle Dautriche, Louise Goupil, Kenny Smith, Hugh Rabagliati

Two-year-olds' eye movements reflect confidence in their understanding of words

Isabelle Dautriche[1,2,*] · Louise Goupil[3,4] · Kenny Smith[5] · Hugh Rabagliati[5]

[1]Laboratoire de Psychologie Cognitive, Aix-Marseille University, CNRS, Marseille, France
[2]Institute of Language, Communication and the Brain, Aix-Marseille University, CNRS, Aix-en-Provence, France
[3]Laboratoire de Psychologie et NeuroCognition, Université Grenoble Alpes, CNRS, Grenoble, France
[4]Department of Psychology, University of East London, London, United Kingdom
[5]School of Philosophy, Psychology Language Sciences, University of Edinburgh, Edinburgh, United Kingdom

[*]To whom correspondence should be addressed. E-mail: isabelle.dautriche@cnrs.fr

Abstract

We study the fundamental issue of whether children evaluate the reliability of their language interpretation, i.e., their confidence in understanding words. In two experiments, two-year-olds (n_1 = 50; n_2 = 60) saw two objects and heard one of them being named; both objects were then hidden behind screens and children were asked to look towards the named object, which was eventually revealed. When children knew the label used, they showed increased post-decision persistence after a correct compared to an incorrect anticipatory look, a marker of confidence in word comprehension (experiment 1). When interacting with an unreliable speaker, children showed accurate word comprehension, but reduced confidence in the accuracy of their own choice, indicating that children's confidence estimates are influenced by social information (experiment 2). Thus, by 2 years, children can estimate their confidence during language comprehension, long before they can talk about their linguistic skills.

Statement of relevance

The capacity to evaluate the reliability of one's own decisions, beliefs and memories, i.e. confidence, is critical in guiding inferential processes in many domains. Whether this capacity develops early in the language domain is far less clear as previous research relied on verbal reports to assess children's ability to explicitly reason about their language understanding. Using a novel paradigm, we provide evidence that the ability to estimate confidence in understanding language is present by at least two years of age and thus, develops in tandem with language comprehension. Our work converges with a growing body of evidence suggesting that monitoring confidence is a fundamental ability that enables humans to actively and adaptively respond to their environment from a very young age and opens critical new questions regarding the role of core metacognition in supporting active and adaptive language learning.

Keywords: language processing; decision confidence; core metacognition; word learning; selective learning; looking-while-listening

Two-year-olds' eye movements reflect confidence in their understanding of words

When we use language to communicate, we are doing more than processing the words we hear; we are trying to infer the speaker's intended meaning given the context we are in (e.g., Clark, 1996; Clayards et al., 2008; Frank and Goodman, 2012; Gibson et al., 2013; Grice, 1975; Levy, 2008; Sperber and Wilson, 1986). Under this account, the ability to estimate confidence, i.e, the likelihood that an interpretation is correct, is thus a central component of language comprehension. A rich body of research now attests that infants and toddlers can recognise words quickly and accurately, much as adults can (e.g., Fernald et al., 2006). But it is far less clear how children develop the capacity to evaluate confidence in these interpretations. Here, we provide novel evidence that toddlers display behavioral markers of confidence in whether they have accurately understood a word, and that their confidence is context sensitive.

The idea that toddlers may be able to estimate their linguistic confidence contrasts with a broader consensus that metalinguistic skills do not develop until quite late (Gleitman and Gleitman, 1979; Hakes, 2012; Levelt et al., 1978). For example, while children know the meanings of many words before their first birthday (Bergelson and Swingley, 2012; Tincoff and Jusczyk, 1999), typically they cannot provide reliable verbal reports on whether a word is familiar, or even whether they know an object's name, until they are about four years old (Marazita and Merriman, 2004).

Though traditional accounts have long assumed that metacognition (i.e., the ability to evaluate one's own cognitive representation) is limited in children below four (e.g., Flavell, 1999), recent research outside the language domain has suggested that certain core aspects of metacognition develop much earlier, long before children can talk about their own cognition (Balcomb and Gerken, 2008; Geurten and Bastin, 2019; Ghetti et al., 2013; Goupil and Kouider, 2019). For example, there is evidence that infants are able to estimate confidence (i.e., the likelihood that a decision is correct (Kepecs et al., 2008; Pouget et al., 2016)) years before they can provide metacognitive verbal reports (Balcomb and Gerken, 2008; Geurten and Bastin, 2019; Goupil and Kouider, 2016; Hembacher and Ghetti, 2014; Kuzyk et al., 2019; Vo et al., 2014). In one recent study Goupil and Kouider (2016) adapted a non-verbal equivalent of the post-decision wagering paradigm (Kepecs et al., 2008), to show that 12-month-old infants can monitor the accuracy of their perceptual decisions. Infants were presented with masked faces that appear for brief durations on the left or right side of a screen, then reappeared a few seconds later as a fully visible reward. Having performed their initial choice (looking either right or left following the prime), infants maintained their gaze longer (i.e. waited longer for the rewarding face) when their initial choice was correct as compared to when it was incorrect. Thus, infants' post-decision persistence primarily varied with the accuracy of their decision, in the absence of any external feedback indexing their performance. This specific pattern of post-decision persistence has been argued to reflect confidence monitoring (i.e., the ability to internally monitor the reliability of one's own decisions), with lower persistence times suggestive of lower confidence in a decision and higher persistence times reflecting greater confidence, a capacity that can also be found in non-human animals (Hampton, 2009; Kepecs et al., 2008; Miyamoto et al., 2017). Post-decision persistence has also been shown to correlate with confidence reports in adults (Kepecs and Mainen, 2012).

These considerations thus raise the possibility that similar behavioral tasks, that do not require a verbal report, might reveal that young children can evaluate their confidence in their understanding of language, for instance in whether they have correctly identified what referent or meaning a speaker intends for a word.

We developed a novel paradigm to assess whether young children's understanding and recognition of words incorporates evaluations of confidence. Our procedure is composed of

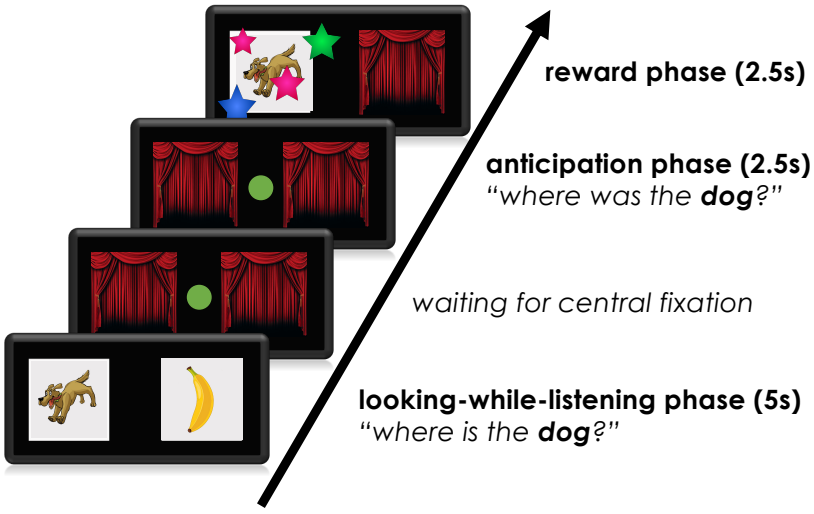

Figure 1. Design of experiment 1. Children's gaze position on a screen was recorded as they completed up to 40 test trials. The figure shows an example of the time course of a test trial where children were tested on the known word "dog".

three distinct phases. First, looking-while-listening. Second, anticipation. Third, reward. The looking-while-listening phase is based on a well-validated eyetracking paradigm of the same name that has frequently been used to assess children's understanding of word meanings (e.g., Bergelson and Swingley, 2012; Fernald et al., 2008; Golinkoff et al., 1987). Participants saw two pictures on a screen and heard one labeled, e.g., "Where is the dog?". To measure children's word recognition accuracy, we recorded their fixations to the named picture over time. The looking-while-listening phase is directly followed by the anticipation phase, a modified anticipatory looking paradigm: the pictures were occluded and participants were asked again to look at the target picture (e.g., "where was the dog?") before it reappeared a few seconds later (the final reward phase). The anticipation task provided a second discrete measure of how the word was understood while the objects were occluded (their first-look decision to look towards the left or right side of the screen in anticipation of the reappearance of the target object), alongside their confidence in that understanding (indexed by post-decision persistence: how long they persisted in gazing toward the hidden object after their first look, in the absence of any further information that could influence their decision). If children can internally evaluate their accuracy in recognizing the target word, then they should show longer persistence times after a correct first-look compared to an incorrect first-look, but only when they actually know the meaning of the word. Critically, the post-decision persistence measure is taken in the absence of any information on the screen about object locations, which ensures that persistence is driven by children's internal evaluation of their first-look accuracy, i.e., their confidence.

Experiment 1

Experiment 1 tested whether children's objective word knowledge modulated their confidence in understanding those words. This is a pre-registered replication of a pilot experiment reported in the SI.

Method

The pre-registration, material, data and the analysis script are available here https://10.17605/OSF.IO/9FAPJ.

Participants. Fifty English-speaking children were included in the final analysis (mean age 23M;8D; SD = 122D, min: 18M;5D, max: 29M;19D; 25 girls). Our sample size was based on Goupil and Kouider (2016)'s Experiment 3. They tested 50 12-month-olds in a post-decision persistence wagering paradigm; A power analysis based on this effect suggested that we should test 70 children to have a power of 80% at the 0.05 alpha level. However, since our participants were older, we limited the number of participants to 50 (pre-registered). An additional 7 children were tested but excluded from the analysis because they did not provide sufficient trials (n = 4; see exclusion criteria below), because their caregiver interfered (n = 1) or because they were born at less than 37 weeks gestational age (n = 1). Participants were recruited in the Edinburgh area.

Procedure and experimental design. Before coming to the lab, parents completed a child vocabulary questionnaire to ensure that they knew the familiar words used in the experiment. During the experiment, children sat on their caregiver's lap in front of a monitor. Caregivers wore opaque glasses, and were asked to not interact with the child during the procedure.

We adapted a version of the post-decision persistence wagering paradigm (see Kepecs et al. 2008 in rats and Goupil and Kouider 2016 in infants) with an anticipation eye-movement paradigm using an eyetracker. The experiment consisted of a series of test trials whose time course is depicted in Figure 1. The trial started with a looking-while-listening phase: Children saw two pictures on the screen depicting either two known objects (*known* word trials; e.g. a dog and a banana) or two unknown objects (*unknown* word trials; e.g., a DNA double helix and a 3D virus shape) and were prompted to look at one of the objects (the target) using its label (e.g., "Where is the dog?" for known words or "Where is the blicket?" for unknown words). The objects were then covered by animated curtains (ending the looking-while-listening phase; 5s including 1s of curtains covering motion). A fixation point (a green circle changing size) then appeared at the centre of the screen between the two curtains and flickered as long as children did not look at it. Once children fixated the fixation point for at least 100ms, the fixation point stopped flickering and the audio started prompting children to find the object labelled during the looking-while-listening phase (e.g. "Did you see the dog?"). The anticipation phase started as soon as children initiated a look towards one of the sides (target curtain; distractor curtain) and lasted for 2.5s of silence with no visual change. The target object then reappeared at the same location as the looking-while-listening phase, along with a rewarding animation and a cheering sound (the reward phase; 2.5s). The reward phase also occurred if the child did not initiate a look in the 4s following the target word offset.

Trials were separated by a 1s pause. No immediately consecutive trials presented the same pictures or words. Target and distractor pictures appeared the same number of times on the right and the left side of the screen. Target side did not repeat more than two times on consecutive trials.

Test trials were presented in blocks of 10, 5 known word trials and 5 unknown word trials. Blocks of trials were repeated as long as children did not show any sign of boredom, to a maximum of 4 repetitions (40 trials). Children received on average 13.16 trials (min: 4; max: 32) after applying the criteria for trial rejection.

The test trials were preceded by two practice trials, designed to familiarise children with the procedure. The first trial consisted of the looking-while-listening phase followed directly by the

reward phase, with no anticipation phase. The second trial included a short anticipation phase of only 500ms.

Materials. Picture stimuli were drawings or photographs of objects on a light gray background. Pictures were always yoked in pairs: 5 pairs for known words (banana/dog, cat/boat, car/bird, shoe/book, hat/ball) and 5 pairs of objects that did not have obvious names in English for unknown words. The familiarisation trials used the pairs star/tree and duck/apple. Parental reports indicated that 7 children did not know one to three of the target known words. Removing these items from the analysis does not change the pattern of results.

For the unknown word trials, 5 novel labels were created: "nurmy", "toma", "blicket", "meb","dax". Each novel label was presented with the same pair of unknown objects across participants. Half of the participants saw the novel label associated with the first object of the pair, and the other half with the second object.

The audio stimuli consisted of one sentence played during the looking-while-listening phase ("Where is the [target]?") and one sentence played just before the anticipation phase ("Where was the [target]?"). All sentences were recorded by a native speaker of English in a child-friendly way.

Criteria for trial and participant exclusion. Trials were rejected if they met any of the following pre-registered rejection criteria: (a) children did not look at either image (target or distractor) for at least 1/2 of the looking-while-listening phase time window (n = 199), (b) The time between the display and the anticipation phase was more than 3 seconds (in order to ensure that the memory of the objects and their location is comparable across trials within and across children; n = 53), (c) Participants did not initiate a look to one of the region of interest (target or distractor) during the anticipation phase (n = 16), (d) this initial look lasted less than 100ms (to avoid implausibly brief responses) (n = 47), and (e) children did not look at either image (target or distractor) for at least 1/2 of the anticipation phase time window (n = 81). These criteria resulted in the removal of 37% of the total number of trials collected.

Participants were excluded if: They had less than 2 trials per word type (known, unknown) after applying the above criteria (a-e), they were premature (born before week 37 of gestation) or they were exposed to less than 50% English input on a weekly basis based on parental estimate.

Measurement and analysis. Gaze position on each trial was recorded via an eye-tracker (Eyelink 1000) with a 2ms sampling rate. All mixed model analyses were performed using the `lme4` package in R (Bates and Sarkar, 2004). For mixed models, we used a maximal random effect structure as supported by the data. P values for main fixed effects are based on likelihood ratio tests, simple effects are reported from the summary table of the model. To check whether the age of children modulated the effects of interest we added Age (in months) as a predictor in all our analyses. Including Age as a predictor was suggested during the review process and was therefore not pre-registered; omitting Age as a predictor does not change other results.

Analysis 1: Word recognition performance.

Recognition during the looking-while-listening phase (pre-registered). We inspected the time course of eye movements during the looking-while-listening phase (5s). We used the proportion of fixations toward the target image as a dependent variable. We conducted three cluster-based permutation analyses (Maris and Oostenveld, 2007) as used previously in eyetracking studies (e.g., Dautriche et al., 2015) on binned data (bins of 50ms excluding away looks) using a custom python script. Word knowledge (known; unknown) was compared to chance by comparing

the average proportion of looks towards the target picture to 50% (the chance level); and one analysis compared the looking proportions between word types.

First-look responses during the anticipation phase (pre-registered). By looking towards one of the sides, toddlers commit to one alternative, which we conceptualize here as a decision, in line with evidence accumulation models of decision making (e.g., Kiani and Shadlen, 2009; Pleskac and Busemeyer, 2010) and many previous studies in children (e.g., Goupil and Kouider, 2016), adults (e.g., Nieuwenhuis et al., 2001) and animals (e.g., Kiani and Shadlen, 2009) and more generally in anticipation designs (e.g., Kovacs and Mehler, 2009).

We modeled participant's first look to one of the hidden objects during the anticipation phase (coded as 1 when the participant initiated a look toward the target and 0 toward the distractor) using a mixed logit model specified as `TargetFirstLook ~ TrialType * Age + (1 | Participant)` with `Age` coded in months and scaled to avoid convergence issues.

Note that first-look responses were, on average, initiated 213ms after target word onset. This is shorter than the minimum latency expected during looking-while-listening tasks (367ms; Swingley and Aslin, 2000; Swingley et al., 1999) but this is expected given the anticipatory nature of the paradigm: when hitting the anticipation phase, children already know the location of the target word, even before hearing the target word label, since they heard it during the immediately preceding looking-while-listening phase (see Figure 1).

Analysis 2: Word recognition confidence.

Persistence times (pre-registered).

To index word recognition confidence we measured persistence times, i.e., how long did participants fixate towards the direction of their first look, during the anticipation phase window. Note that we did not have any a priori hypothesis about how persistence might vary overall between the different stimuli (known vs. unknown objects/words), rather we focused on whether persistence times, within each stimulus type, depend upon first look accuracy. We used the following mixed model: `Persistence ~ TrialType * FirstLook * Age + (1 | Participant)`. Note that we had to simplify the pre-registered analysis which included a random slope for `TrialType * FirstLook` as the model turned out to be singular. Persistence times were log transformed in order to respect the assumption of normality (non-transformed data are used for display purposes in the figures). `Age` was coded in months and scaled to avoid convergence issues. We did not include random effects for items as our number of items was low, but additional analyses did not reveal important item-specific variation. In particular there were no significant differences in persistence times among test trials that feature pairs of inanimate objects (e.g., shoe/book) vs. pairs that feature one animate and one inanimate objects (e.g., banana/dog).

Gaze-shift frequency (post-hoc).

It was suggested during the review process that we also conduct a gaze-shift frequency analysis (GSF). Several studies have reported that GSF, how frequently participants saccade between options presented on the screen, reflects explicitly-reported confidence (Folke et al., 2016; Sepulveda et al., 2020): adult participants who shift their gaze more often between visually-presented options were more likely to report lower confidence in their choice than participants who shifted their gaze less (see also in children Leckey et al. 2020). We analyzed GSF during the anticipation phase. We expected children to switch gaze between options more frequently when they were less confident in their response (i.e., in the unknown word trials) compared to when they were confident in their response (i.e., in the known word trials). Gaze-shift frequency was calculated as the number of times participants shifted their gaze from one area of interest (Target position; Distractor position) to the other during the anticipation phase. We used the following mixed-model: `GSF ~ TrialType * Age + (1 | Participant)`.

Results

Analysis 1: Word recognition performance

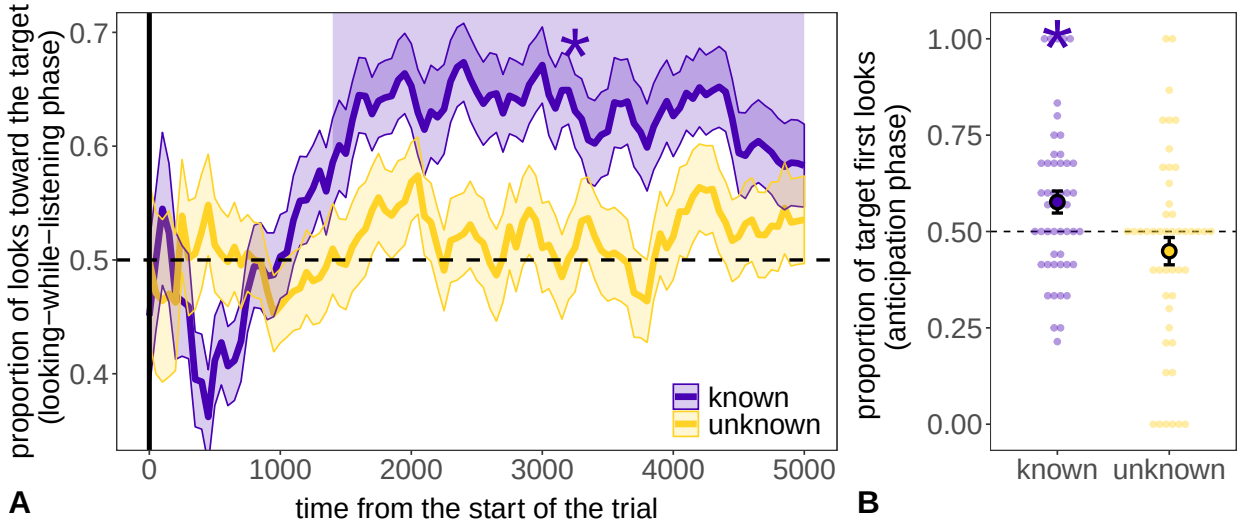

Figure 2. Word recognition performance (experiment 1). (A) Mean proportion of target looks during the looking-while-listening phase for known words (purple) and unknown words (yellow). The purple shaded area represent the time range where the proportion of target looks for the known words was significantly above the chance level (0.5). The ribbon surrounding each curve represents the standard error of the mean obtained at each time bin for each condition. **(B)** Mean proportion of target first-look during the anticipation phase depending on word knowledge (known; unknown). Error bars represent standard errors of the mean. Dots represent individual means.

Recognition during the looking-while-listening phase (Figure 2A). During the looking-while-listening phase, children hearing known words looked toward the target significantly above chance (from 1400 ms to the end of the trial, $p < .001$), and as expected, did not show any preference for the target object when hearing unknown words ($p > .3$). There was a significant difference between known and unknown words (from 2100 ms to 3350 ms; $p = .006$). Further analyses revealed that this effect was more robust in the older participants than in the younger ones (see details in the SI).

First-look responses during anticipation phase (Figure 2B). There was a main effect of trial type (known word vs. unknown words; $\chi^2(1) = 7.49$, $p = .006$). For known words, children were significantly more likely than chance to initiate a first look toward the hidden target ($M = 0.58, SE = 0.03$, $\beta = 0.11, z = 2.45, p = .01$) but not for unknown words ($M = 0.45, SE = 0.03$, $\beta = -0.15, z = -1.22, p = .22$). There was also a main effect of age ($\chi^2(1) = 6.58$, $p = .01$) with older children more likely to initiate a first look towards the hidden target than younger children (see figure in the SI). No other effects nor interaction were significant.

Additional analyses reported in the SI show that first-looks responses were more robust when children took more time to initiate their first-look responses for known words while response latencies did not affect accuracy for unknown words (see details in the SI). This suggests that first-look decisions are a mixture of informed responses in which children fully process the target word and retrieve the most probable location of the target referent (longer latencies), as well as

early responses initiated before being able to fully process the target word, i.e., potential mistakes reflecting a variety of additional factors (stimulus preference or stimulus complexity, side biases, etc)

Analysis 2: Word recognition confidence

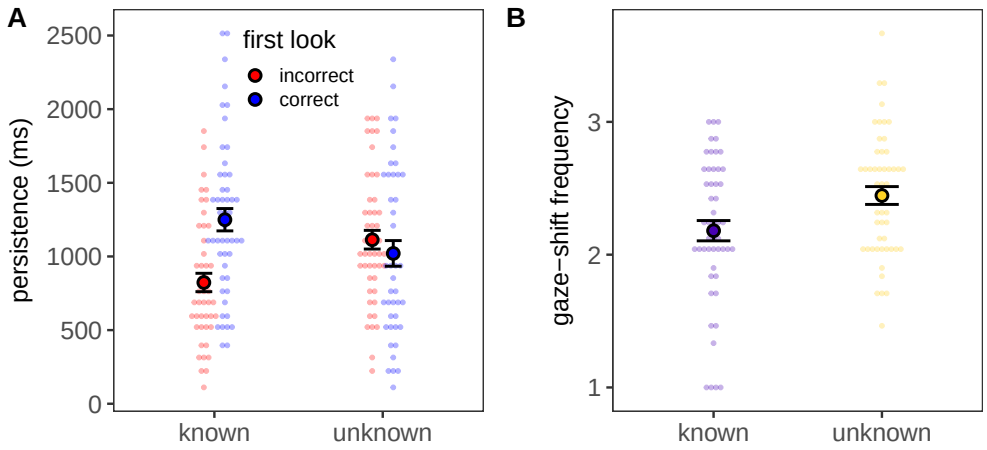

Figure 3. Word recognition confidence (experiment 1). (A) Relationship between persistence times and first-look accuracy depending on word knowledge (known; unknown). Persistence times were averaged separately for correct (blue) and incorrect (red) first looks for each level of word knowledge. **(B)** Mean number of gaze switches between area of interests (target, distractor) during the anticipation phase. Error bars represent standard errors of the mean. Dots represent individual means.

Persistence times (Figure 3A). Persistence times during the anticipation phase were affected by first-look accuracy ($\chi^2(1) = 4.30$, $p = .04$), with participants looking longer after a correct compared to an incorrect first look. Yet this pattern depended on whether the words were known or not, as would be expected if persistence indexes the confidence associated with children's decisions about what each word meant. When tested on known words, participants showed longer persistence times after making a correct first look as compared to an incorrect first look ($M_{correct} = 1228ms$, $SE_{correct} = 78ms$, $M_{incorrect} = 823ms$, $SE_{incorrect} = 62ms$, $\beta = 0.33, t = 3.92, p < .001$) but accuracy did not affect persistence when tested on unknown words ($M_{correct} = 1061ms$; $SE_{correct} = 98ms$, $M_{incorrect} = 1145ms$, $SE_{incorrect} = 70ms$, $\beta = -0.07, t = -0.85, p = .39$), and the interaction between word knowledge and first look accuracy was significant ($\chi^2(1) = 11.207$, $p < .001$).

There was also a significant interaction between age and first look ($\chi^2(1) = 8.14$, $p = .004$) indicating that older children were more likely than younger children to display higher persistence after a correct first look than after an incorrect first look. No other main effects or interactions were significant.

Four further analyses provided evidence consistent with this key finding reflecting confidence, and inconsistent with lower-level counter-explanations. First, if this pattern of persistence reflects confidence, then we may expect the difference between correct and incorrect looks to be larger on trials where participants trade off speed for accuracy, echoing the rich literature showing a correlation between performance and confidence in human adults (see for review Fleming and Lau, 2014). And consistent with this, post-decision persistence differed between correct and

incorrect responses for both slow and fast responses (see related analysis of accuracy above) and this difference was stronger for slow responses (see details in the SI).

Second, although the effect of accuracy was larger for slow response times (see SI), we did not find any evidence of a simple correlation between latency to first look and persistence times (p = .8), which rules out the possibility that children's persistence times can be explained by a low-level association between persistence times and response times (see details in SI).

Third, we also did not find any evidence that first look responses (correct vs. incorrect) reflected different degrees of word-referent knowledge activation. For instance it could be that children who initiated an incorrect first look did so because they knew the tested word less well than children who initiated a correct first look (despite parental reports being similar) or were just less motivated by the task and thus did not reactivate their word-referent knowledge as well as motivated children. Yet, children's target looking behavior during the looking-while-listening phase (which reflects their word-referent knowledge as well as their motivation for the task) was not different between those trials that lead to a correct vs. an incorrect first look (no cluster found; see details in the SI).

Finally we did not find evidence that children were more likely to persist for longer towards the object they had favored during the looking-while-listening phase (see details in SI). This suggests that persistence times do not simply reflect the after-effects of low-level attentional processes operating during the looking-while-listening phase (i.e., side or object).

In sum, because children did not receive external feedback indexing their performances during the anticipation phase, the difference in persistence times suggests that they were using internal evidence to evaluate whether or not they had made the correct decision, i.e., monitoring the confidence associated with their understanding of the words. This effect was most visible by two years of age, where participants showed a reliable look-to-target performance for known words during the looking-while-listening task and on first-looks during the anticipation phase. This could be the result of weaker language processing skills in younger children, or weaker confidence monitoring, or both.

Gaze-shift frequency (post-hoc; Figure 3B). The number of gaze shifts during the anticipation phase was modulated by word knowledge. Children shifted their gaze more often when tested on unknown words (M = 2.45, SE = 0.07) than when tested on known words (M = 2.18, SE = 0.08, $\chi^2(1)$ = 4.08, p = .04). No other effect or interaction was significant. This is suggestive of lower confidence, with children actively shifting their gaze between the two possible options when they know that they don't know (Folke et al., 2016; Leckey et al., 2020).

Our results thus show that two-year-olds can monitor their word recognition performance in a word recognition task.

Experiment 2

While Experiment 1 showed that persistence times index children's confidence about what a word is likely to refer to, Experiment 2 aimed to establish that these confidence estimates reflect a child's confidence that they understand what a word is intended to mean.

Our method draws on evidence that, by age two, children can account for speakers who use words idiosyncratically, like labeling a ball as "dog". If an unreliable, idiosyncratic speaker teaches a two-year-old a new word, e.g., that a novel object is called a "wug", then that child will restrict the domain of that word to that specific individual, and will not generalize its use with other individuals (Dautriche et al., 2021; Koenig and Woodward, 2010). This suggests that the reliability

of a speaker may impact children's confidence in how words are used even when children show similar accuracy levels. To wit, if an unreliable speaker tells the child to "look at the cat" on a trial in which both a cat and a boat are hidden, then the child may infer that "cat" probably refers to the cat, as a best guess. But they may not be confident in that response, because the speaker has been unreliable in the past, and would thus show a reduced difference between post-decision persistence times following correct vs. incorrect responses.

In Experiment 2, two-year-olds first watched a video in which a confederate demonstrated themselves to be either a reliable or unreliable speaker, and then taught the child two new words. Then, participants completed a word recognition task as in Experiment 1, in which the same speaker used a combination of familiar words and the newly-taught novel words. For both novel and known words, we predicted that children would show accurate recognition, but with lower confidence when the speaker is unreliable (Figure 4).

Method

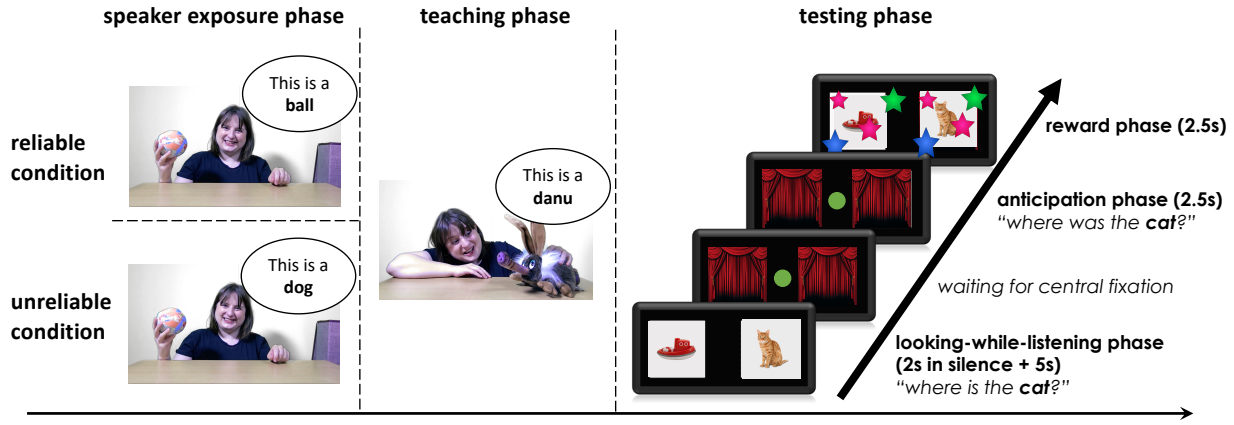

Figure 4. Design of experiment 2. The experiment consisted of 3 phases: 1) the speaker exposure phase where a speaker labeled familiar objects either using correct labels (e.g., calling a ball a "ball"; the reliable condition) or incorrect labels (e.g., calling a ball a "dog"; the unreliable condition); 2) the teaching phase where the speaker taught two novel words ("danu" and "modi") for two novel objects, and the testing phase, similar to Experiment 1, which tested recognition and confidence in both known words (as pictured; different from the labels used during the exposure phase) and novel words (with the two novel objects displayed on the screen). The test trials used the same speaker as the exposure phase.

Participants. Sixty English-speaking children were included in the final analysis, 30 in the reliable condition (mean age 30M;19D, SD = 53D, min: 27M;26D; max: 34M;14D, 12 boys) and 30 in the unreliable condition (mean age 29M;28D, SD = 80D, min: 24M;14D, max: 35M;29D, 14 boys). We tested on average older children than in Experiment 1 because past literature using a similar design mostly focused on older children (a single study tested under-two-year-old children Luchkina et al. 2018). The number of participants was estimated using experiment 1's data on the results of the first 16 trials and by considering the experiment as a between-subject design. A power analysis based on this effect suggests that we should test at least 40 children per condition to have a power of 80% at the 0.05 alpha level. Since we tested children that are on average older than in experiment 1, we decided to limit the number of participants to 30 per condition

(pre-registered). An additional 20 children were tested but excluded from the analysis because they did not provide sufficient trials (n = 11; see exclusion criteria below), because they did not want to participate in the experiment (n = 5), because of sibling or caregivers interference (n = 3) or because of technical issues (n = 1). Participants were recruited in the Edinburgh area.

Procedure, experimental design and material The experiment was composed of 3 phases as described in Figure 4:

Speaker exposure phase. Participants saw a video of a native English female speaker playing with five objects and labelling them. Each object was taken out of a box individually, labelled three times and put back into the box. The same five objects were used across the two conditions: a tiger puppet, a banana, a ball, a shoe and glasses. In the reliable condition, the speaker used the correct label to refer to the objects. In the unreliable condition, the speaker used incorrect labels that did not refer to any other objects seen in the video (flower, car, dog, book, star).

Teaching phase. Participants saw two 30s videos, each teaching them one novel word. In each video the speaker (of Phase 1) showed a novel object and labelled it five times using one of two novel words ("danu" or "modi"). The novel objects were two unfamiliar animals (see pictures in SI).

Testing phase. Test trials matched the procedure of Experiment 1 (see Figure 1), but used new audio stimuli recorded by the reliable/unreliable speaker. We implemented two changes to the trial time course. First, the looking-while-listening phase started with the simultaneous presentation of the two pictures in silence (2s), in order to increase children's performance during the looking-while-listening phase by giving them sufficient time to explore the picture before hearing the target word. Second, both pictures reappeared on the screen during the reward phase. This was done in order to maintain the unreliability of the speaker for children in the unreliable condition, but was implemented in both conditions. Importantly, this did not impact children's motivation to look at the target object during the reward phase (see details in the SI).

The testing phase was composed of 16 test trials: 8 known words trials and 8 novel words trials. The known trials used 8 objects that did not appear during the Speaker exposure phase (orange/butterfly, spoon/duck, cat/boat, hat/fish). According to parental reports, 5 children did not know 1 to 3 of these known words. Removing these items from the analysis does not change the pattern of results. Each pair was shown twice, and each referent named once. The novel trials showed the two newly-learned objects, with each being named four times. The smaller number of trials in this study matched the average number of trials completed in Experiment 1.

Criteria for trial and participant exclusion. Same as in Experiment 1. This removed 43% of the total number of trials collected.

Analyses. We conducted the same analyses as in Experiment 1. Since neither `Age` nor any of its interactions with other predictors were significant we removed it from the model. Note that in our pre-registration, we discarded the first-look analysis as a measure of word knowledge as it seemed to be too noisy following Experiment 1 and previous research showing that first-looks are a more variable index of word understanding than fixation proportions in the looking-while-listening paradigm (Naigles and Gelman, 1995; Naigles, 1996). However, for the sake of completeness, we report these results here.

Preliminary results Since there was no learning difference between the specific novel word being tested ("danu" vs. "modi"; p_{min} = .20 using a cluster-based permutation analysis on the

proportion of target looks during the looking-while-listening phase), we compared participants' behaviour across conditions (reliable vs. unreliable) collapsing looking behaviour for all trials testing novel words. As in experiment 1, the animacy of the target did not significantly affect participants' persistence times.

Results

Analysis 1: Word recognition performance

Recognition during the looking-while-listening phase (Figures 5A and 5C). As in Experiment 1, the looking-while-listening phase of the test trials showed that children readily recognised known words. They looked toward the target significantly above chance in the reliable condition (from 600ms to 4700ms, $p < .001$) and in the unreliable condition (from 550ms to 4350ms, $p < .001$), with no difference between these conditions. For the newly-taught novel words, we observed a similar pattern: children looked toward the target significantly above chance in both conditions (reliable: from 850ms to 2050ms, $p = .007$, and from 2450ms to 3600ms, $p = .001$; unreliable: from 2900ms to 3550ms, $p = .036$), again, with no difference between conditions.

First-look responses during the anticipation phase (post-hoc; Figures 5B and 5D). Overall, participants looked towards the target above chance when tested on known words ($\beta = 0.30, z = 2.58, p = .01$). They were significantly more likely than chance to initiate a first look toward the target in the unreliable condition ($M = 0.60, SE = 0.03$; $\beta = 0.39, z = 2.34, p = .02$). Performance was not significantly above chance in the reliable condition ($M = 0.55, SE = 0.03$, $\beta = 0.21, z = 1.29, p = .20$), possibly because in this condition the average latency of first-looks was 174ms, lower than in the unreliable condition (200ms) or that the latency of the known word condition of experiment 1 (213ms), and thus may include more early, pre-emptive looks than the other condition (see details in the SI). However, there was no difference between conditions ($\chi^2(1) = 0.09$, $p = .76$). For novel words, participants were not more likely than chance to look at the target in either the reliable ($M = 0.52, SE = 0.05$, $\beta = 0.03, z = 0.17, p = .86$) or the unreliable ($M = 0.55, SE = 0.04$, $\beta = 0.14, z = 0.75, p = .45$) conditions.

As a whole, our results show that children recognize familiar words when tested by both a reliable or an unreliable speaker. Their display-phase responses also show that children learned the novel words in both conditions, replicating previous studies (Koenig and Woodward, 2010). Following experiment 1, the first-look accuracy was not high (as it may not always correspond to a response selection, see SI), but critically was comparable across conditions for both word types, allowing us to analyze how word recognition confidence may vary across conditions while controlling for accuracy.

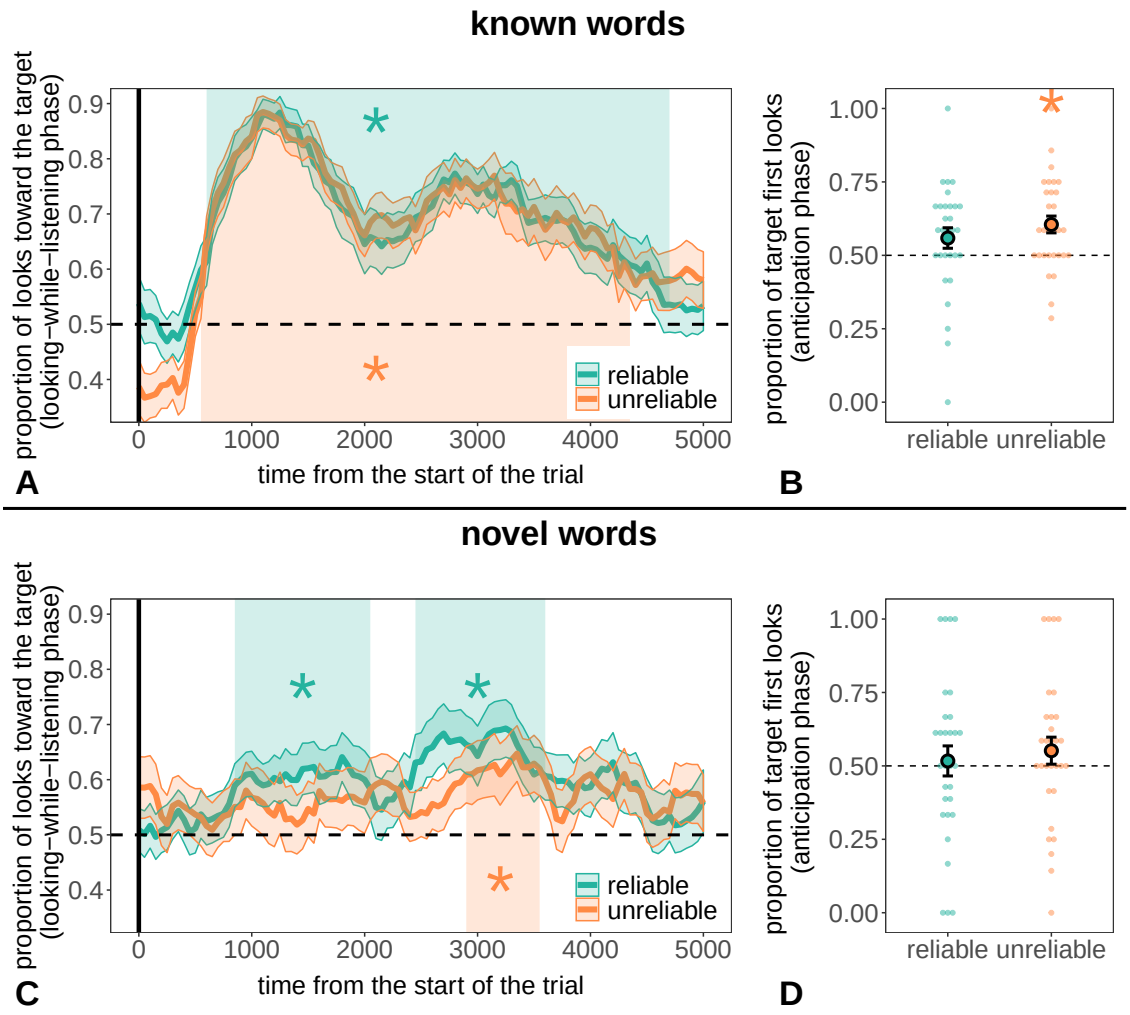

Figure 5. Word recognition performance (experiment 2). For known **(A)** and novel **(C)** words: Proportion of looks towards the target picture, time-locked to the beginning of the target word for the reliable condition (in green) and for the unreliable condition (in orange). The ribbon surrounding each curve represents the standard error of the mean obtained at each time bin for each condition. Children looked to the target significantly above chance (0.5) in the reliable condition (light green shaded area) and in the unreliable condition (light orange shaded area). For known **(B)** and novel **(D)** words: Mean first-look accuracy during the anticipation phase in the reliable condition (green) and in the unreliable condition (orange). The dots represent individual data points and the error bars standard error of the mean.

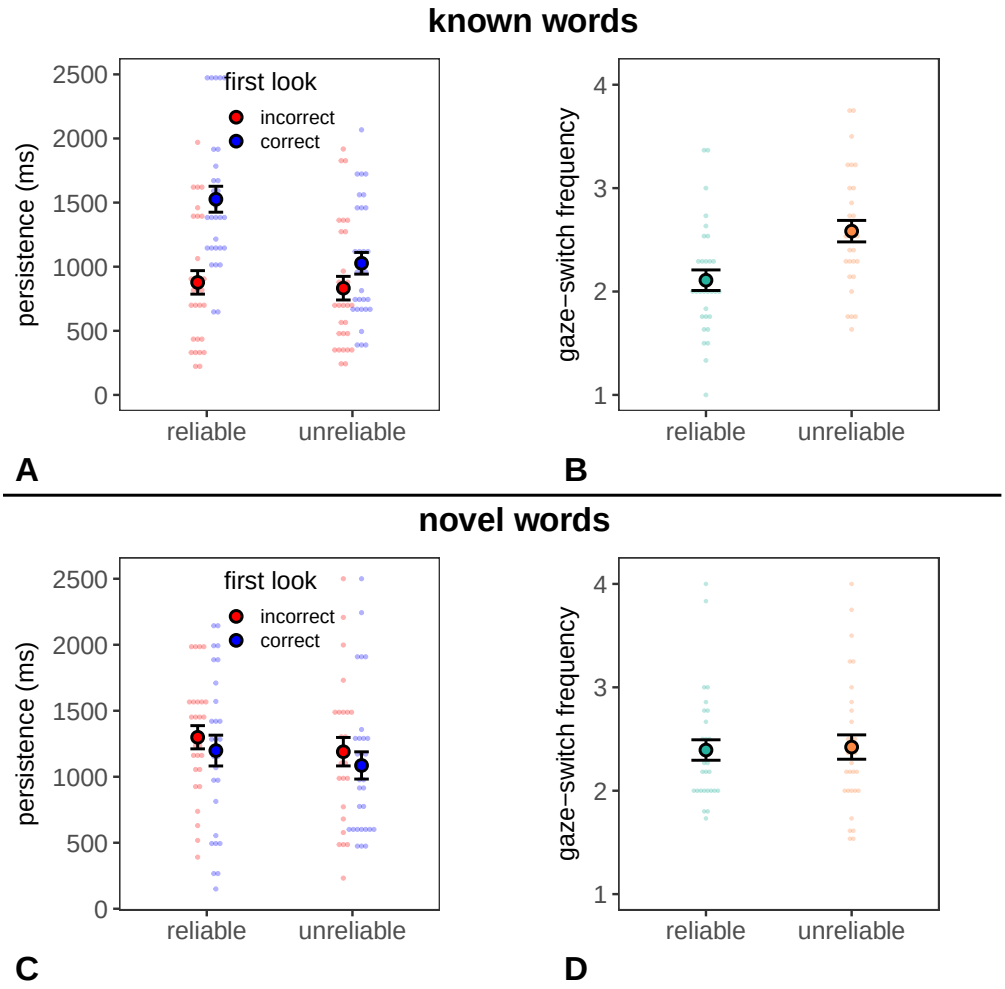

Figure 6. Word recognition confidence (experiment 2). For known **(A)** and novel **(C)** words: Relationship between persistence times and first-look accuracy depending on condition (reliable; unreliable). Persistence times were averaged separately for correct (blue) and incorrect (red) first look for each condition. For known **(B)** and novel **(D)** words: Mean number of gaze switches between area of interests (target, distractor) during the anticipation phase. Error bars represent standard errors of the mean. Dots represent individual means.

Analysis 2: Word recognition confidence

Persistence times (Figures 6A and 6C) Following our pre-registered plan we analyzed persistence times separately for known and novel words (a combined analysis can be found in the SI). For known words, children's persistence was not only influenced by their first-look accuracy, but also by the reliability of the speaker, leading to a significant interaction between these factors ($\chi^2(1) = 4.24$, $p = .04$). Overall, children persisted longer after a correct first look than an incorrect first look (main effect of accuracy, $\chi^2(1) = 20.35$, $p < .001$) but they did so more when the speaker was reliable rather than unreliable. For the reliable speaker, persistence times were significantly longer after correct rather than incorrect first looks ($M_{correct} = 1526ms$, $SE_{correct} = 101ms$, $M_{incorrect} = 878ms$, $SE_{incorrect} = 92ms$, $\beta = 0.542, t = 4.70, p < .001$), while for the unreliable speaker this difference was marginally significant ($M_{correct} = 985ms$, $SE_{correct} = 85ms$, $M_{incorrect} = 813ms$, $SE_{incorrect} = 90ms$, $\beta = 0.205, t = 1.76, p = .08$). The main effect of speaker reliability on persistence times was marginal ($\chi^2(1) = 3.56$, $p = .06$).

For the novel words, however, persistence times were not modulated by either first-look accuracy or condition (all ps > 0.11), despite children having shown that they could recognise these novel words during the looking-while-listening phase. This suggests that children were able to recognise the referents of the novel words, but that they were not yet confident in their lexical decisions (at least as indexed by persistence times), or that they could not yet evaluate their confidence, presumably because the words were newly-learned.

Further analyses again ruled out lower-level counterexplanations. First, as in Experiment 1, there was no evidence that children's persistence times can be explained solely by low-level associations between persistence times and response times, nor between persistence times and object preference during the looking-while-listening phase (see SI).

Moreover, analysis of behaviour during the reward phase indicated that participants' memory for the target words, and target word locations, were matched across the reliable and unreliable speaker conditions. Specifically, children in both conditions looked at the target above chance, with no difference between conditions, even though linguistic stimuli were absent. Thus, the idiosyncrasy of the speaker did not affect memory reinstatement processes. This also rules out the possibility that persistence times may index children's confidence in remembering the location of the object rather than their linguistic confidence, as memory is unaffected by speaker reliability (see more details in the SI).

Gaze-shift frequency (post-hoc; Figures 6B and 6D). We tested whether the number of gaze shifts during the anticipation phase was influenced by speaker reliability and by word type and the interaction between these two factors. Children shifted their gaze more when the speaker was unreliable compared to when she was reliable ($\chi^2(1) = 3.99$, $p = .05$). This was significant for known words ($M_{reliable} = 2.10$, $SE_{reliable} = 0.09$, $M_{unreliable} = 2.69$, $SE_{unreliable} = 0.13$, $\beta = 0.54, t = 3.19, p = .002$) but not for unknown words ($M_{reliable} = 2.39$, $SE_{reliable} = 0.10$, $M_{unreliable} = 2.49$, $SE_{unreliable} = 0.14$, $\beta = 0.29, t = 0.23, p = .79$). The interaction between word type and speaker condition was significant ($\chi^2(1) = 8.13$, $p = .004$). This result is congruent with the post-decision persistence analysis, and confirms that children in the unreliable condition were less confident than children in the reliable condition, in particular when tested on known words. There was no main effect of word type (($\chi^2(1) = 0.10$, $p = .75$) but, in the reliable condition, children generated more gaze shifts when tested on novel words than when tested on known words ($\beta = 0.27, t = 2.25, p = .02$) suggestive of greater uncertainty in novel word trials in the reliable condition.

Our results thus show that children's confidence estimates are influenced by social information.

General Discussion

These two experiments show that, by 24 months, children's looking behavior reveals their confidence in understanding a word: they persist more in recognition decisions when they have reasons to be sure about a word's meaning.

Critically, because children's confidence appeared to be affected by the reliability of the speaker, this suggests that children were evaluating not only what the words they heard meant, but what they thought the speaker intended the words to mean. This is important because it is consistent with pragmatic accounts of language comprehension (Clark, 1996; Grice, 1975; Sperber and Wilson, 1986) as well as with modern noisy-channel models of adult language processing (e.g., Clayards et al., 2008; Levy, 2008), which highlight that sentence comprehension involves both decoding the current signal, and integrating that signal with prior knowledge about what meanings a speaker

is likely to express, in order to derive the most probable interpretation. Children's context-relative confidence estimates suggest that they can already integrate their processing of a signal with their prior knowledge of a speaker (e.g., the speaker's reliability), and thus implies that, by age two, they are already able to process words and sentences using an active, noisy-channel strategy.

The ability to estimate confidence during language comprehension could play an important role throughout language development. For instance, confidence estimates could be used by children to optimise how they allocate attention during learning (e.g., attending to situations in which they have low-confidence in their interpretation of words, see Zettersten and Saffran (2019)). Moreover, confidence estimates could also guide children's interrogative behaviors: low confidence in having understood a word would be a signal for children to request clarification from their caregivers, either behaviorally or through verbal requests (Bazhydai et al., 2020; Butler et al., 2020; Hembacher et al., 2020; Jimenez et al., 2018). Our findings also suggest that confidence estimates are responsive to the social context. Specifically children exposed to an idiosyncratic speaker who used labels unconventionally show reduced confidence in their interpretations. Such re-calibration of confidence may have important implications for learning: under-confidence may lead children to be overly receptive to any additional information they may receive about a word meaning; over-confidence, on the other hand, could make them indifferent to it (Rollwage and Fleming, 2021; Rollwage et al., 2020).

Throughout this discussion, we have interpreted our participants' eye-movement behaviour in terms of decision accuracy (for the first location of eye movements) and confidence (for the persistence following initial choices). But might these data also be accounted for by simpler mechanisms? For instance, it has been previously suggested that simpler interpretations in terms of first-order processes such as attention or memory could explain both decision accuracy and post-decision persistence in similar paradigms (Carruthers, 2020; Gliga and Southgate, 2016). These alternative interpretations, however, find little support in our results. First, and most importantly, our results show for the first time that young children's post-decision persistence can be dissociated from their ability to perform a task, and varies depending on the social context, as is the case in adults (Jacquot et al., 2015). Such a dissociation would not be expected if accuracy and persistence were driven by a single mechanism. Second, neither participants' memory for the target words, target word locations nor attention during the display phase predicted persistence patterns. This taken as a whole represents the strongest evidence to date that young children's post-decision persistence truly reflect confidence, rather than performance, attention or memory. It may be that confidence directly reflects properties of the decision-making process (e.g., the distance between accumulated evidence and a decision bound, or an evaluation of decision time) (Kiani and Shadlen, 2009; Pereira et al., 2021). Alternatively, it could be that confidence reflects core metacognitive monitoring even in young children (Goupil and Kouider, 2019). The finding that speaker idiosyncrasy can differentially impact first-look accuracy and post-decision persistence favours this latter interpretation.

Finally, our results show that one of the most widely-used methods in infant language research, the looking-while-listening paradigm, can elide very different states of label-referent understanding. For instance, Experiment 2 found highly similar looking-while-listening performance for recognising known words uttered by reliable versus unreliable speakers, but our persistence measure revealed differences in confidence levels. We suggest that our paradigm could be an important new tool for more precisely evaluating the interpretations infants give to words and sentences, especially to assess children's emerging pragmatic skills.

In sum, our work converges with a growing body of evidence suggesting that monitoring confidence is a fundamental ability that enables humans to actively and adaptively respond to their environment from a very young age (Ghetti et al., 2013; Goupil and Kouider, 2019). It extends

previous results by showing that toddlers' capacity for confidence monitoring is not restricted to the evaluation of simple perceptual decisions, but extends to socially-informed conventional knowledge. The influence that monitoring confidence has on early lexical development is currently unknown, but we hope that these results will stimulate interest in characterizing the role that confidence monitoring plays in supporting active and adaptive language learning.

Authors contribution

ID developed the study concept. All authors contributed to the study design. Testing, data collection and analysis were performed by ID. All authors contributed to the writing of the manuscript.

Acknowledgments

Many thanks to Jessica Brough, Jenny Chim, Anna Hall, Rachel Kindellan and Rebekah Oakley for data collection and to the actor and voice of our stimuli, Emma Healey. The research leading to this results has received funding from the ESRC (ES/N017404/1 and ES/N005635/1).

References

Balcomb, F. K. and Gerken, L. (2008). Three-year-old children can access their own memory to guide responses on a visual matching task. *Developmental Science*, 11(5):750–760.

Bates, D. and Sarkar, D. (2004). *lme4 library*. Accessed.

Bazhydai, M., Westermann, G., and Parise, E. (2020). "i don't know but i know who to ask": 12-month-olds actively seek information from knowledgeable adults. *Developmental science*, 23(5):e12938.

Bergelson, E. and Swingley, D. (2012). At 6-9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, 109(9):3253–3258.

Butler, L. P., Ronfard, S., and Corriveau, K. H. (2020). *The questioning child: Insights from psychology and education*. Cambridge University Press.

Carruthers, P. (2020). *Questions in development*. The questioning child, Cambridge, UK.

Clark, H. H. (1996). *Using language*. Cambridge university press.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., and Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, 108(3):804–809.

Dautriche, I., Goupil, L., Smith, K., and Rabagliati, H. (2021). Knowing how you know: Toddlers reevaluate words learned from an unreliable speaker. *Open Mind*, 5:1–19.

Dautriche, I., Swingley, D., and Christophe, A. (2015). Learning novel phonological neighbors: Syntactic category matters. *Cognition*, 143:77–86.

Fernald, A., Perfors, A., and Marchman, V. A. (2006). Picking up speed in understanding: Speech processing efficiency and vocabulary growth across the 2nd year. *Developmental psychology*, 42(1):98.

Fernald, A., Zangl, R., Portillo, A. L., and Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language. *Developmental psycholinguistics: On-line methods in children's language processing*, pages 113–132.

Flavell, J. H. (1999). Cognitive development: Children's knowledge about the mind. *Annual review of psychology*, 50(1):21–45.

Fleming, S. M. and Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, 8.

Folke, T., Jacobsen, C., Fleming, S. M., and De Martino, B. (2016). Explicit representation of confidence informs future value-based decisions. *Nature Human Behaviour*, 1(1):0002.

Frank, M. C. and Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998.

Geurten, M. and Bastin, C. (2019). Behaviors speak louder than explicit reports: Implicit metacognition in 2.5-year-old children. *Developmental science*, 22(2):e12742.

Ghetti, S., Hembacher, E., and Coughlin, C. A. (2013). Feeling uncertain and acting on it during the preschool years: A metacognitive approach. *Child Development Perspectives*, 7(3):160–165.

Gibson, E., Bergen, L., and Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110(20):8051–8056.

Gleitman, H. and Gleitman, L. (1979). Language use and language judgment. In *Individual differences in language ability and language behavior*, pages 103–126. Elsevier.

Gliga, T. and Southgate, V. (2016). Metacognition: Pre-verbal infants adapt their behaviour to their knowledge states. *Current Biology*, 26(22).

Golinkoff, R. M., Hirsh-Pasek, K., Cauley, K. M., and Gordon, L. (1987). The eyes have it: Lexical and syntactic comprehension in a new paradigm. *Journal of child language*, 14(1):23–45.

Goupil, L. and Kouider, S. (2016). Behavioral and Neural Indices of Metacognitive Sensitivity in Preverbal Infants. *Current Biology*, 26(22):3038–3045.

Goupil, L. and Kouider, S. (2019). Developing a reflective mind: From core metacognition to explicit self-reflection. *Current Directions in Psychological Science*, 28(4).

Grice, H. (1975). In Cole, P. and Morgan, J. L., editors, *Syntax and Semantics, Vol. 3, Speech Acts*, pages 41–58. Academic Press, New York.

Hakes, D. T. (2012). *The development of metalinguistic abilities in children*, volume 9. Springer.

Hampton, R. R. (2009). Multiple demonstrations of metacognition in nonhumans: Converging evidence or multiple mechanisms? *Comparative cognition & behavior reviews*, 4:17.

Hembacher, E., deMayo, B., and Frank, M. C. (2020). Children's social information seeking is sensitive to referential ambiguity. *Child development*, 91(6):e1178–e1193.

Hembacher, E. and Ghetti, S. (2014). Don't look at my answer: Subjective uncertainty underlies preschoolers' exclusion of their least accurate memories. *Psychological science*, 25(1768).

Jacquot, A., Eskenazi, T., Sales-Wuillemin, E., Montalan, B., Proust, J., Grèzes, J., and Conty, L. (2015). Source unreliability decreases but does not cancel the impact of social information on metacognitive evaluations. *Frontiers in Psychology*, 6.

Jimenez, S., Sun, Y., and Saylor, M. M. (2018). The process of active word learning. In *Active Learning from Infancy to Childhood*, pages 75–93. Springer.

Kepecs, A. and Mainen, Z. F. (2012). A computational framework for the study of confidence in humans and animals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1322–1337.

Kepecs, A., Uchida, N., Zariwala, H. A., and Mainen, Z. F. (2008). Neural correlates, computation and behavioural impact of decision confidence. *Nature*, 455(7210):227.

Kiani, R. and Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *science*, 324(5928):759–764.

Koenig, M. A. and Woodward, A. L. (2010). Sensitivity of 24-month-olds to the prior inaccuracy of the source: Possible mechanisms. *Developmental Psychology*, 46(4):815–826.

Kovacs, A. M. and Mehler, J. (2009). Flexible Learning of Multiple Speech Structures in Bilingual Infants. *Science*, 325(5940):611–612.

Kuzyk, O., Grossman, S., and Poulin-Dubois, D. (2019). Knowing who knows: Metacognitive and causal learning abilities guide infants' selective social learning. *Developmental science*, page e12904.

Leckey, S., Selmeczy, D., Kazemi, A., Johnson, E. G., Hembacher, E., and Ghetti, S. (2020). Response latencies and eye gaze provide insight on how toddlers gather evidence under uncertainty. *Nature Human Behaviour*, 4(9):928–936.

Levelt, W. J., Sinclair, A., and Jarvella, R. J. (1978). Causes and functions of linguistic awareness in language acquisition: Some introductory remarks. In *The child's conception of language*, pages 1–14. Springer.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Luchkina, E., Sobel, D. M., and Morgan, J. L. (2018). Eighteen-month-olds selectively generalize words from accurate speakers to novel contexts. *Developmental Science*, 21(6):e12663.

Marazita, J. M. and Merriman, W. E. (2004). Young children's judgment of whether they know names for objects: The metalinguistic ability it reflects and the processes it involves. *Journal of Memory and Language*, 51(3):458–472.

Maris, E. and Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of neuroscience methods*, 164(1):177–190.

Miyamoto, K., Osada, T., Setsuie, R., Takeda, M., Tamura, K., Adachi, Y., and Miyashita, Y. (2017). Causal neural network of metamemory for retrospection in primates. *Science*, 355(6321):188–193.

Naigles, L. G. and Gelman, S. A. (1995). Overextensions in comprehension and production revisited: Preferential-looking in a study of dog, cat, and cow. *Journal of Child Language*, 22(1):19–46.

Naigles, L. R. (1996). The use of multiple frames in verb learning via syntactic bootstrapping. *Cognition*, 58(2):221–251.

Nieuwenhuis, S., Ridderinkhof, K. R., Blom, J., Band, G. P., and Kok, A. (2001). Error-related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task. *Psychophysiology*, 38(5):752–760.

Pereira, M., Megevand, P., Tan, M. X., Chang, W., Wang, S., Rezai, A., Seeck, M., Corniola, M., Momjian, S., Bernasconi, F., et al. (2021). Evidence accumulation relates to perceptual consciousness and monitoring. *Nature communications*, 12(1):1–11.

Pleskac, T. J. and Busemeyer, J. R. (2010). Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychological review*, 117(3):864.

Pouget, A., Drugowitsch, J., and Kepecs, A. (2016). Confidence and certainty: distinct probabilistic quantities for different goals. *Nature neuroscience*, 19(3):366.

Rollwage, M. and Fleming, S. M. (2021). Confirmation bias is adaptive when coupled with efficient metacognition. *Philosophical Transactions of the Royal Society B*, 376(1822):20200131.

Rollwage, M., Loosen, A., Hauser, T. U., Moran, R., Dolan, R. J., and Fleming, S. M. (2020). Confidence drives a neural confirmation bias. *Nature communications*, 11(1):1–11.

Sepulveda, P., Usher, M., Davies, N., Benson, A. A., Ortoleva, P., and De Martino, B. (2020). Visual attention modulates the integration of goal-relevant evidence and not value. *Elife*, 9:e60705.

Sperber, D. and Wilson, D. (1986). *Relevance: Communication and cognition*, volume 142. Harvard University Press Cambridge, MA.

Swingley, D. and Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, 76(2):147–166.

Swingley, D., Pinto, J. P., and Fernald, A. (1999). Continuous processing in word recognition at 24 months. *Cognition*, 71(2):73–108.

Tincoff, R. and Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds.

Psychological science, 10(2):172–175.

Vo, V., Li, R., and Kornell, N., P. A. C. J. (2014). Young children bet on their numerical skills: Metacognition in the numerical domain. *Psychological science*, 25(1712).

Zettersten, M. and Saffran, J. R. (2019). Sampling to learn words: Adults and children sample words that reduce referential ambiguity. *Developmental Science*, page e13064.