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]STMS UMR9912 (IRCAM/CNRS/Sorbonne Université), Paris, France

[4]University of East London, London, UK

[5]University of Edinburgh, Edinburgh, UK

Abstract

We study the fundamental issue of whether children assess the reliability of their interpretation while processing language, i.e., their confidence in understanding words. In two experiments, two-year-olds (n = 50 and n = 60) were asked to look toward one of two objects which were labelled then hidden behind screens. When children knew the label used, they showed increased persistence in their initial choice after a correct compared to an incorrect response, a marker of decision confidence in word recognition (experiment 1). When interacting with an unreliable speaker (experiment 2), children showed accurate word recognition, but reduced confidence in the accuracy of their own choice, as indexed by post-decision persistence, indicating that children monitor their confidence in what words are intended to mean. These results provide the first evidence that toddlers can estimate their confidence in their word recognition decisions, long before they can explicitly reflect upon and talk about their linguistic understanding.

Statement of relevance

The capacity to represent the reliability of one's own decisions, 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 talk about their language understanding. Using a novel implicit paradigm, we provide evidence that the ability to estimate confidence in language understanding 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 metacognition in supporting active and adaptive language learning.

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

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

A central question for theories of human development concerns how we learn to understand spoken language. By adulthood, words are understood through a combination of automatic and controlled processes (Meyer et al., 2007). During a conversation, for example, automatic processes cause us to quickly recognise sounds, and retrieve associated word meanings, while controlled, mindreading and metacognitive abilities allow us to reflect on exactly what the speaker intended to mean, and to compute a degree of confidence in our interpretations. A rich body of research now attests that the automatic processes of word recognition emerge early in development: Infants and toddlers can recognise words quickly and accurately, much as adults can (e.g., Fernald et al., 2006). But whether the controlled, metacognitive capacity also develops early is far less clear. Here, we provide novel evidence that that is the case, showing that even toddlers are able to estimate their confidence in whether they have accurately understood a word.

The adult capacity for metacognitive monitoring extends across domains: we can assess our confidence in our decisions, our percepts, our beliefs and our memories (Dunlosky and Metcalfe, 2008; Mamassian, 2016; Yeung and Summerfield, 2012). This ability to estimate certainty is particularly important for modern models of language processing, which assume that adult listeners commit to the most likely interpretation of a word or sentence by weighting different interpretative options according to their reliability (e.g., Clayards et al., 2008; Frank and Goodman, 2012; Gibson et al., 2013; Levy, 2008). In children, however, metacognitive reasoning about language has typically been considered a late-emerging skill. For example, children cannot accurately judge 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) and there is no evidence that children can judge whether a sentence is grammatical before age five (Ambridge et al., 2008). Why these metacognitive skills emerge so late in the linguistic domain is unclear, but could be a consequence of the particular tasks that have been used, which typically rely upon explicit reports, often recorded verbally, and which thus may place insoluble demands on children's limited capacity for talking about language.

If so, implicit tasks, that do not require a verbal metacognitive report, could reveal that even infants and toddlers have a rudimentary metacognitive competence in the linguistic domain.

Indeed, recent methodological advances have provided evidence that basic forms of metacognition, such as the capacity to estimate *decision confidence* (hereafter "confidence")–the likelihood that a decision is correct (Kepecs et al., 2008; Pouget et al., 2016)– are present in infants and toddlers in non-linguistic domains (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 appearing for brief durations on the left or right side of a screen, that 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. Importantly, this dependency of post-decision persistence on decision accuracy vanished when decisions were made on invisible faces, which shows that it was dependent upon participants being aware of the stimuli, a core functional feature of subjective confidence that has also been documented in adult populations (Persaud, 2007). This specific pattern of post-decision persistence has been argued to reflect performance monitoring, or metacognitive sensitivity (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 (Lak et al., 2014)) and can also be found in non-verbal species (Hampton, 2009; Kepecs et al., 2008; Miyamoto et al., 2017).

These considerations thus raise the possibility that young children may also be able to 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. Such a demonstration of early implicit metalinguistic evaluations would not only be important for theories of word recognition, but also for theories of metacognition: It would provide evidence that early metacognition is domain general, such that children can evaluate not only perceptual decisions, but also socially-informed conventional knowledge.

We developed a novel paradigm to assess whether young children's understanding and

recognition of words incorporates evaluations of confidence. Our method integrates the above-described measure of post-decision persistence (Goupil and Kouider, 2016; Kepecs et al., 2008) into a looking-while-listening procedure, a well-validated eyetracking task 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) (Figure 1). 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 during this display phase. Then, 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. This latter task provided a second discrete measure of how the word was understood (their first-look decision), 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 indexing their performances). 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.

In Experiment 1, we tested whether children's objective word knowledge (confirmed by parental report) modulated their confidence in understanding these words. In Experiment 2, we tested whether social information about the reliability of a speaker impacted confidence. The results of these experiments show that by two years of age, children can monitor the confidence associated with their language understanding, long before they can explicitly reflect upon and talk about their linguistic understanding.

Experiment 1

Method

The pre-registration, material, data and the analysis script are available here https://osf.io/9fapj/?view_only=36d8222b32f049c497dc38efcd987776. Pilot results are reported in the SI.

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). The number of participants was

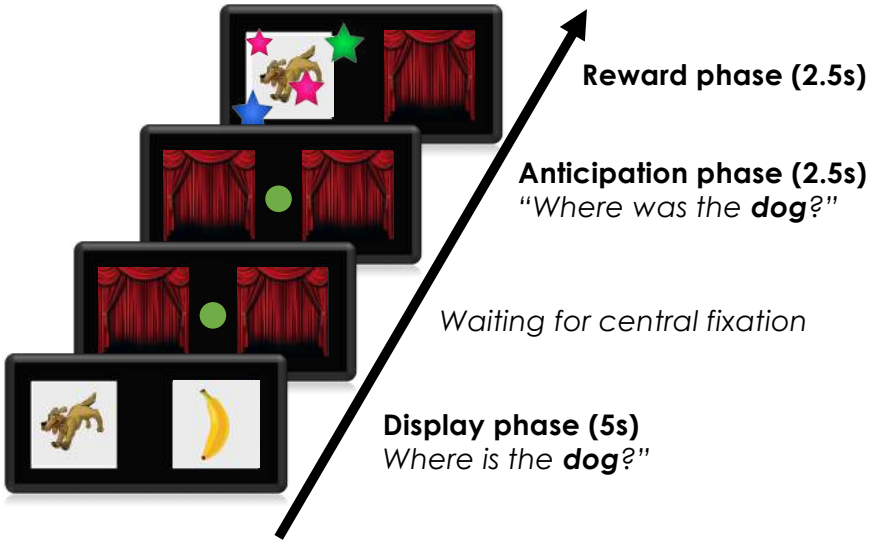

Figure 1. Design of experiment 1. Children's gaze position on a screen was recorded as they completed up to 40 test trials. The figure shows an example of the time course of a test trial where children were tested on the known word "dog".

estimated by using Goupil and Kouider (2016) (experiment 3)'s data who tested 50 participants in a post-decision persistence wagering paradigm similar to ours in 12-month-olds. A power analysis based on their effect suggests that we should test 70 children to have a power of 80% at the 0.05 alpha level. Since we tested older children, we decided to limit the number of participants to 50. 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 SA (n = 1). Participants were recruited in the XXXX 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. Children first 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 object (the target) using its label (e.g., "Do you see the dog?" for known words or "Do you see the blicket?" for unknown words). The objects were then covered by animated curtains (ending the display 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 display 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 (in silence). If children did not initiate a look in the 4s following the target word offset, the trial continued normally. The target object then reappeared at the same location it was seen during the display period with a rewarding animation and a cheering sound (the reward phase; 2.5s).

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.

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.

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 display phase ("Do you see the [target]?") and one sentence played just before the anticipation phase ("Did you see 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 complying with any of the following pre-registered criteria: (a) We obtained less than 50% of eye-data during the display phase, (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), (c) Participants did not initiate a look to one of the region of interest (target or distractor) during the anticipation phase, (d) this initial look lasted less than 100ms (to avoid implausibly fast responses), and (d) we obtained more than 50% of eye-data during the anticipation phase. This removed 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-d), they were premature (born before 37 of gestation) or they were exposed to less than 60% 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. For analysing the time course of eye movement, we used a cluster-based permutation analysis (Maris and Oostenveld, 2007) implemented in a custom python script. All remaining 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. Details of the models can be found in the SI and in the online script.

Results

Analysis 1: Word recognition performance

Recognition during the display phase (Figure 2A). During the display phase, children hearing known words looked toward the target significantly above chance (from 1400 ms to the end of the trial, $p < .001$, tested via a cluster-based permutation analysis (Maris and Oostenveld, 2007)), 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; $p = .006$).

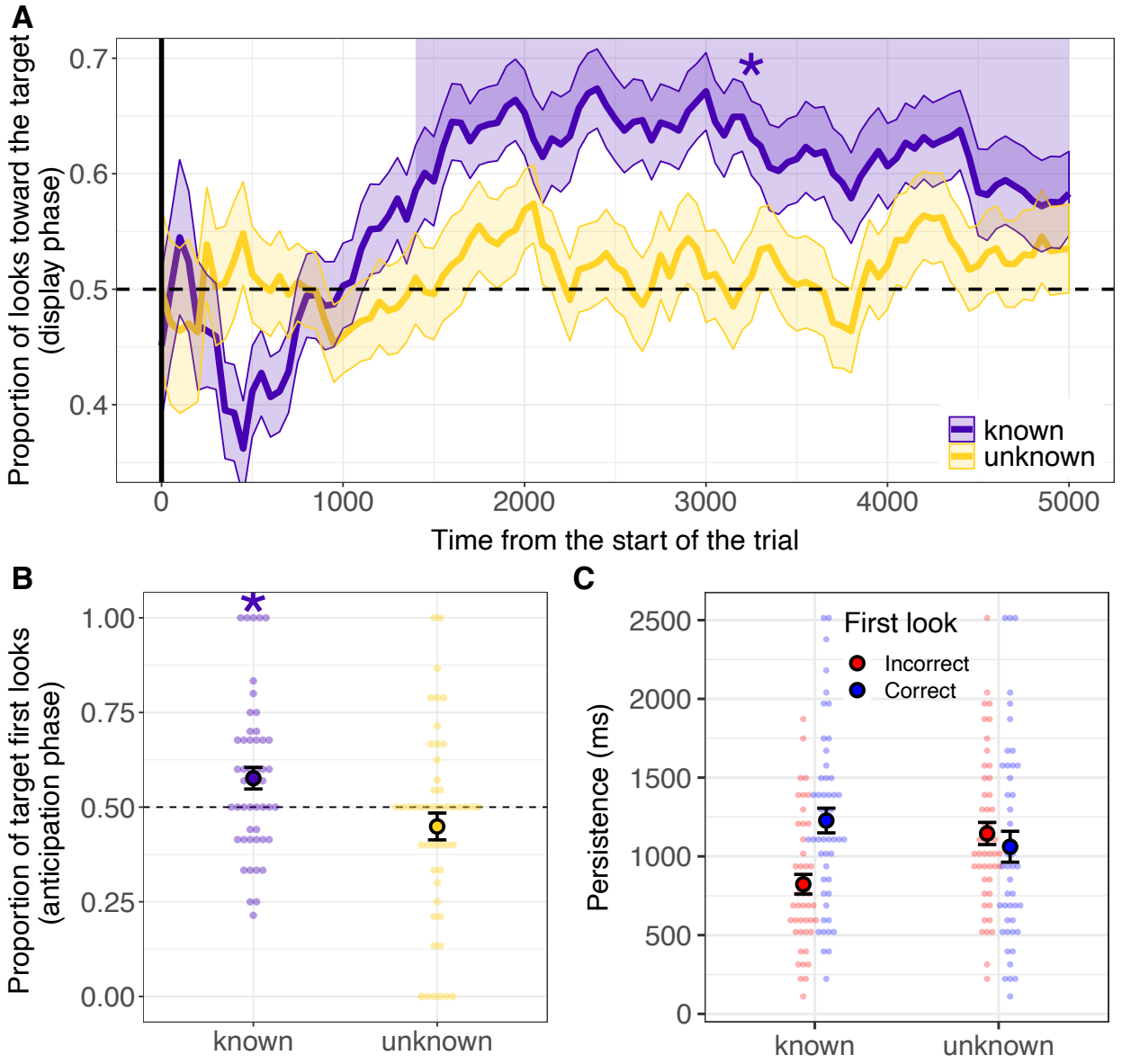

Figure 2. Results of experiment 1. (A) Mean proportion of target looks during the display phase for known words (purple) and unknown words (yellow). The purple shaded area represent the time range where the proportion of target looks for the known words was significantly above the chance level (0.5). 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). **(C)** 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. Error bars represent standard errors of the mean. Dots represent individual mean persistence times.

First-look responses (Figure 2B). For known words, children were significantly more likely than chance to initiate a first look toward the target ($M = 0.58; SE = 0.03; \beta = 0.27; z = 2.14; p = .03$) but not for unknown words ($M = 0.45; SE = 0.03; \beta = -0.16; z = -1.25; p = .21$). The difference in target first

look proportion between known and unknown words was significant ($\beta = -0.43; z = -2.69; p = .007$). Importantly, the average first-look response was initiated 149ms (SD= 670ms) after target word onset, which is considerably shorter than the standard response latency expected in a looking-while-listening procedure (367ms after target word onset, although it is often more see Swingley, 2012). This suggests that first looks are a mixture of responses in which children fully process the target word and retrieve the location of the target referent, as well as responses initiated before being able to fully process the target word, i.e., potential mistakes, that the child would be able to correct if they can monitor the accuracy on their behaviour without external evidence.

Analysis 2: Word recognition confidence (Figure 2C)

Persistence times during the anticipation phase were differentially affected by first look accuracy, depending 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 ($\beta = 0.32; t = 3.83; p < .001$) but accuracy did not affect persistence when tested on unknown words ($\beta = -0.06; t = -0.68; p = .50$), and the interaction between these factors was significant ($\chi^2(2) = 9.97;\ p = .002$). Notably, when participants were tested on known words, then they persisted less after an incorrect first look than after a correct or incorrect first look when tested on unknown words ($\beta_{correct} = -0.19; t = -2.1; p = .04$; $\beta_{incorrect} = -0.26; t = -3.09; p = .002$) while there was no difference between their persistence after a correct first look on known words and their persistence times on unknown words ($\beta_{correct} = 0.11; t = 1.38; p = .17$; $\beta_{incorrect} = 0.06; t = 0.81; p = .41$). This suggests that the effect was mostly driven by participants detecting their errors in the known word condition. There was also a main effect of first-look accuracy ($\chi^2(2) = 4.52;\ p = .03$) and no main effect of word type ($p = .20$). We did not find any evidence of an effect of response times (the time taken by participants to initiate their first look) on persistence times ($p = .8$) ruling 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).

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

word recognition task. Their persistence times, an implicit measure of confidence, were shorter when they had incorrectly chosen the location of a hidden referent, but only when they knew the meaning of the tested word. Importantly, 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.

Experiment 2

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 (Clark, 1996; Grice, 1975; Sperber and Wilson, 1986). While Experiment 1 showed that persistence times index children's confidence about what a word means, in Experiment 2 we 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 (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.

Importantly, such an effect would also rule out a lower-level alternative hypothesis of Experiment 1, namely that children simply persist for longer when they first look towards the location they had also favored during the display phase. In particular, no effect of reliability would be expected if persistence times index the child's confidence in remembering the correct location of the referent (which is unaffected by speaker reliability), but an effect should be present if

persistence indexes their confidence that they know what the speaker meant. Thus, 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 3).

Method

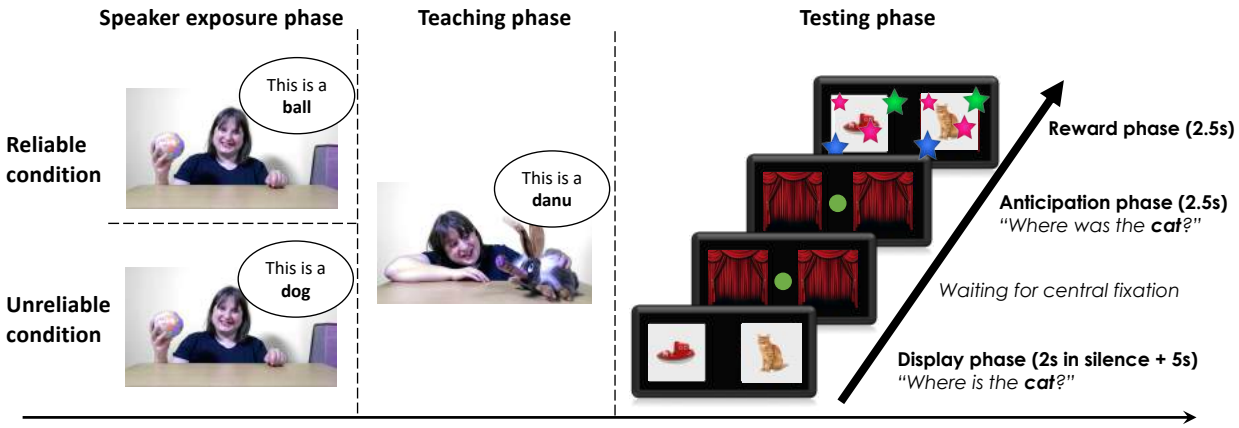

Figure 3. Design of experiment 2. The experiment consisted of 3 phases: 1) the speaker exposure phase where a speaker labeled familiar objects either using correct labels (e.g., calling a ball a "ball"; the reliable condition) or incorrect labels (e.g., calling a ball a "dog"; the unreliable condition); 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 for two reasons: First, we did not observe any qualitative difference between the younger and the older children on their persistence score (see SI) and second, 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. 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 XXXX area.

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

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 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 display phase started with the simultaneous presentation of the two pictures in silence (2s), in order to increase children's performance during the display 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.

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). 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. Since we did not expect any learning difference between the specific novel word being tested ("danu" or "modi"), we compared participants' behaviour across conditions (reliable vs. unreliable) collapsing looking behaviour for all trials testing novel words. Details of the analyses can be found in the SI and in the online script for analysis.

Results

Analysis 1: Word Recognition performance

Recognition during the display phase (Figures 4A and 4D). As in Experiment 1, the display 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 (Figures 4B and 4E). Overall, participants looked above chance to the 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; β = 0.21; z = 1.29; p = .19); however, there was no difference between conditions (β = 0.18; z = 0.76; p = .44). 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; β = 0.03; z = 0.17; p = .86) or the unreliable (M = 0.55, SE = 0.04; β = 0.14; z = 0.75; p = .40) 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, 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.

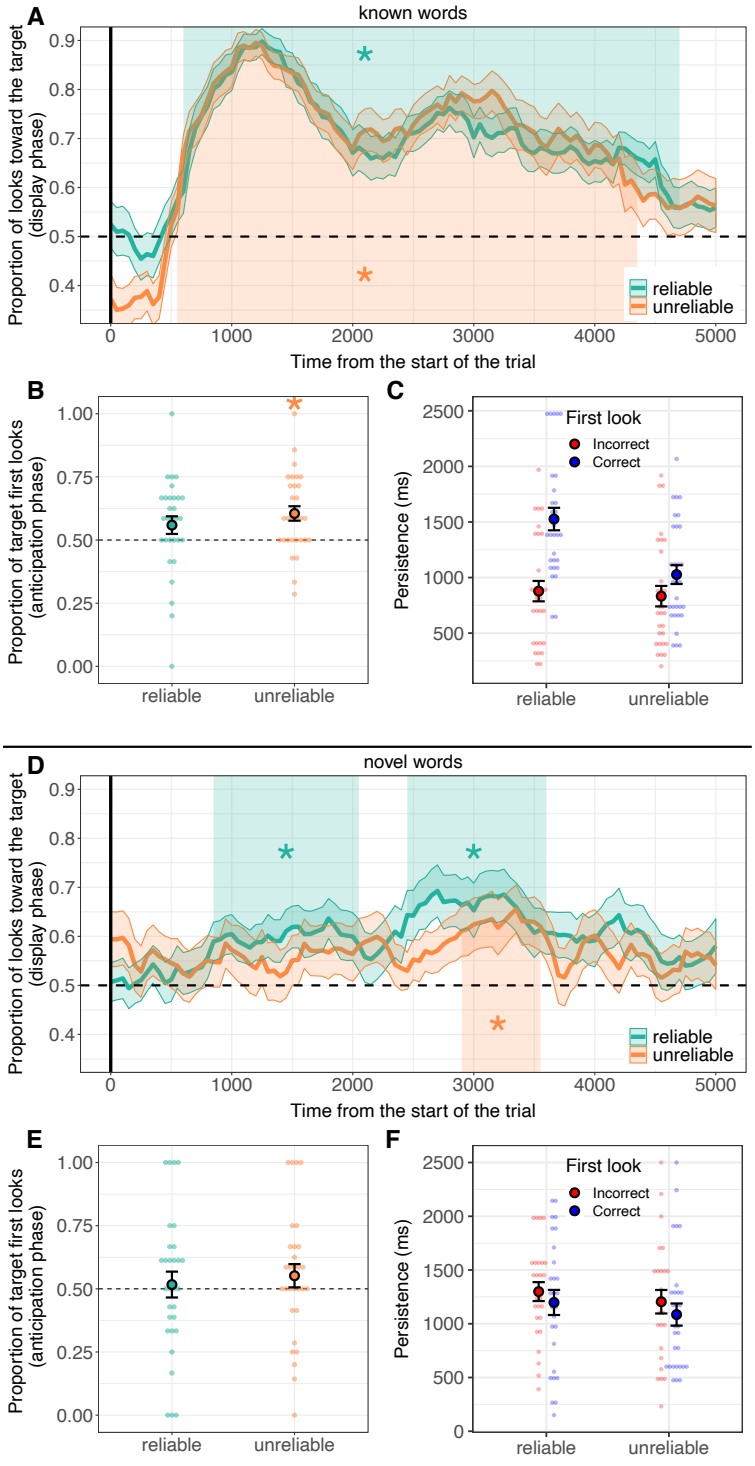

Figure 4. Results of experiment 2. For known **(A)** and novel **(D)** words: Proportion of looks towards the target picture, time-locked to the beginning of the target word for the reliable condition (in green) and for the unreliable condition (in orange). The ribbon surrounding each curve represents 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 **(E)** words: Mean first-look accuracy during the anticipation phase in the reliable condition (green) and in the unreliable condition (orange).For known **(C)** and novel **(F)** 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. The dots represent individual data points and the error bars standard error of the mean.

Analysis 2: Word recognition confidence (Figures 4C and 4F)

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 ($\chi^2(2) = 4.24$; $p = .04$). Overall, children persisted longer after a correct first look than an incorrect first look (main effect of accuracy, $\chi^2(2) = 20.35$; $p < .001$) 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 look ($\beta = 0.542; t = 4.70; p < .001$), while for the unreliable speaker this difference was marginally significant ($\beta = 0.205; t = 1.76; p = .08$). The main effect of speaker reliability on persistence times was marginal ($\chi^2(2) = 3.56$; $p = .06$). 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 display phase.

Our results show that children's confidence estimates are influenced by social information. Specifically, for known words, speaker reliability did not affect children's recognition accuracy, yet it did affect their confidence: When the speaker was reliable, children persisted for longer after making an accurate decision, but when the speaker was unreliable, accuracy had less effect on persistence. Such an effect of speaker reliability 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 should be unaffected by speaker reliability. An effect of speaker reliability on confidence estimates was not visible on novel words: Regardless of speaker reliability, children showed accurate recognition (at least during the display phase), but their accuracy did not affect their persistence times. 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 assess their confidence, presumably because the words were newly-learned.

General Discussion

These two experiments show that, by 24 months, children's looking behavior reveals their decision confidence in how they have understood a word: they persist more in recognition decisions when

they have reasons to be sure about a word's meaning. Our data are thus the first to establish that two-year-old children estimate their confidence in their language understanding, at least implicitly, long before they become able to explicitly talk about their language knowledge, in the fourth or fifth year of life.

Critically, because children's confidence appeared to be derived relative to 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 theories that provide a high-level 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.

Beyond language processing, our results also have implications for theories of early metacognition, showing for the first time that young children's confidence can be dissociated from their ability to perform a task, and varies depending on the social context. This critically represents the strongest evidence to date that implicit measures of confidence, such as post-decision persistence, truly reflect metacognitive process, rather than performance (Gliga and Southgate, 2016) or affective states (Carruthers, 2016). By contrast, this dissociation would be expected if children's post-decision persistence stems from a metacognitive, inferential process integrating both information linked to decision making, as well as contextual factors that may be relevant for determining the reliability of one's own knowledge, as is the case in adults (Jacquot et al., 2015).

Finally, our results highlight a methodological lacuna in one of the most widely-used methods in infant language research: the looking-while-listening paradigm. As we showed this method 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.

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). 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 metacognition plays in supporting active and adaptive language learning.

References

Ambridge, B., Pine, J. M., Rowland, C. F., and Young, C. R. (2008). The effect of verb semantic class and verb frequency (entrenchment) on children's and adults' graded judgements of argument-structure overgeneralization errors. *Cognition*, 106(1):87–129.

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.

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.

Carruthers, P. (2016). Are epistemic emotions metacognitive? *Philosophical Psychology*, 30(1-2).

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.

Dunlosky, J. and Metcalfe, J. (2008). *Metacognition*. Sage Publications.

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.