Subjective confidence influences word learning in a cross-situational statistical learning task

Citation for published version:
Dautriche, I, Rabagliati, H & Smith, K 2021, 'Subjective confidence influences word learning in a cross-situational statistical learning task', *Journal of Memory and Language*, vol. 121, 104277. https://doi.org/10.1016/j.jml.2021.104277

Digital Object Identifier (DOI):
10.1016/j.jml.2021.104277

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Journal of Memory and Language

Subjective confidence influences word learning in a cross-situational statistical learning task

Isabelle Dautriche[1,2] · Hugh Rabagliati[3] · Kenny Smith[3]

[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]University of Edinburgh, Edinburgh, United Kingdom

Word count: 8393

Author Note: The pre-registration (experiment 3 only), the data and the script for their analysis (all experiments) are available here: https://osf.io/upndk/.

Acknowledgments

This research received funding from the ESRC under the Future Research Leaders schemes (ES/N017404/1 and ES/N005635/1).

Abstract

Learning is often accompanied by a subjective sense of confidence in one's knowledge, a feeling of knowing what you know and how well you know it. Subjective confidence has been shown to guide learning in other domains, but has received little attention so far in the word learning literature. Across three word learning experiments, we investigated whether and how a sense of confidence in having acquired a word meaning influences the word learning process itself. First, we show evidence for a confirmation bias during word learning in a cross-situational statistical learning task: Learners who are highly confident they know the meaning of a word are more likely to persist in their belief than learners who are not, even after observing objective evidence disconfirming their belief. Second, we show that subjective confidence in a word-meaning modulates inferential processes based on that word, affecting learning over the whole lexicon: Learners who hold high confidence in a word-meaning are more likely to use that word to make mutual exclusivity inferences about the meaning of other words. We conclude that confidence influences word learning by modulating both information selection processes and inferential processes and discuss the implications of these results for word learning models.

Keywords: word learning, cross-situational learning, confidence, metacognition

Learning the meaning of a word, even the simplest labels, is a hard problem (Quine, 1960). Even when a competent French speaker uses the word "chat" to refer to their furry pet, a French-learning listener may be clueless. "Chat" could indeed refers to cats, but it could also refer to the sofa the cat is sleeping on, another pet situated nearby, or even some property of the cat such as its furriness, cuteness and so on. Yet, despite this ubiquitous referential uncertainty, infants and adults only need a few exposures to a word to home in on its meaning (Bloom, 2002; Carey & Bartlett, 1978; Xu & Tenenbaum, 2007). An important body of experimental work provides evidence that word learners can reduce referential uncertainty in the moment, by exploiting linguistic (e.g., L. Gleitman, 1990), social (e.g., Baldwin, 1993) and attentional (e.g., L. B. Smith & Samuelson, 2006) cues, but also by accumulating evidence across individually ambiguous exposures, a process called cross-situational statistical learning (Siskind, 1996; K. Smith, Smith, & Blythe, 2011; L. Smith & Yu, 2008; Yu, Smith, Klein, & Shiffrin, 2007). For example, the word "chat" may be used more often when a cat is present than when a dog is present, and thus, over time, such co-occurrences regularities would support the right meaning for the word "chat" over others.

An important question concerns how hypotheses about a word meaning are formed and evaluated across learning exposures. Some evidence supports an associative learning mechanism in which learners track the entire system of word-meaning co-occurences across learning instances such that the meaning of a word is the referent with the strongest statistical correlation over all learning instances (Fazly, Alishahi, & Stevenson, 2010; S. Frank, Goldwater, & Keller, 2009; Siskind, 1996; Yu et al., 2007). Other evidence supports an hypothesis-testing mechanism in which learning is more discrete, with learners selecting the most likely meaning for a word in a given moment and subsequently confirming or falsifying this hypothesis as new information becomes available in subsequent word usages (Medina, Snedeker, Trueswell, & Gleitman, 2011; Trueswell, Medina, Hafri, & Gleitman, 2013; Yang, 2020). While the use of one or the other mechanism may depend on attentional and memory demands (Yurovsky & Frank, 2015), both mechanisms focus on how learners use their *objective* experience with the world, in and across learning exposures, to generate and evaluate word meaning hypotheses and do not attempt to capture the influence of more *subjective* processes.

During learning broadly speaking, people not only gather information from the external world but also consult and evaluate their subjective confidence in that information, i.e., the degree to which they believe their current knowledge or decision is correct. In memory tasks, subjective confidence is thought to guide self-regulated learning (Bjork, Dunlosky, & Kornell, 2013; Metcalfe & Finn, 2008). For instance, people tend to spend more time studying items that are judged to be more difficult (for a review see Son & Metcalfe, 2000). Subjective confidence predicts adaptive behavior in decision making tasks: when confidence is low, people are likely to sample more information before giving a response (Desender, Boldt, & Yeung, 2018) and are more likely to change their mind were the same choice to be presented again (Folke, Jacobsen, Fleming, & De Martino, 2016). Such empirical evidence supports the idea that subjective confidence is important for learning from mistakes, for generating predictions, and for guiding subsequent learning or decision-making, even in the absence of feedback (see also Meyniel, Sigman, & Mainen, 2015; Yeung & Summerfield, 2012). However, the role that subjective confidence may play during word learning, in particular in how hypotheses for a word meaning are evaluated, has received little attention so far. Here we investigate whether and how one's subjective confidence influences the word learning process itself.

Previous work suggests that adult learners do indeed track the uncertainty associated with their word knowledge in a cross-situational statistical learning task. In these tasks, participants are asked to learn the meanings of several words across a series of trials that simulate the ambiguity of the real world (e.g., Yu & Smith, 2007). Each trial is individually ambiguous as it consists of one or more words and several candidate referents; each such exposure is therefore consistent with multiple possible mappings from words to referents, but as the number of observations increases, participants can successfully determine the correct referent for each word, by using the cross-situational statistics of word-referent co-occurrences. Yurovsky and Yu (2008) measured participants' subjective confidence after each trial in such a task, and found that participants' confidence is positively correlated with their accuracy, with this correlation increasing across learning trials as objective uncertainty reduces. This echoes numerous results in decision-making tasks: When the objective information clearly favors one option, accuracy and confidence will both be high; when the information is ambiguous then both accuracy and confidence will be low (for review see Fleming & Lau, 2014). Thus confidence and performance are distinct but well-correlated estimates of the objective uncertainty present in the world (Fleming & Daw, 2017; Kiani & Shadlen, 2009). Critically this suggests that to demonstrate that confidence has an independent effect on word learning, one needs to show that confidence predicts performance, i.e., referent selection, while controlling for the objective uncertainty of the environment.

We present three experiments with adult learners showing that subjective confidence influences 1) referent selection in a cross-situational word learning task and 2) mutual exclusivity-based inferences about the meaning of novel words. Across all experiments participants were presented with a series of word learning trials for each to-be-learned word. On each trial, one word was presented along with four possible referents, and participants were prompted to select a referent as their guess for the word's meaning; we recorded participants' referent choice and their subjective confidence in their choice. In experiment 1, we manipulated the order of exposures to construct a series of learning trials that, if participants are influenced by their own subjective confidence rather than purely relying on the objective properties of their input, should lead participants to display a confirmation bias (Nickerson, 1998), i.e. a tendency to ignore disconfirming information when their confidence is high. We show that holding high confidence in one's knowledge of a word meaning modulates subsequent information processing: learners who were more confident that they knew a word's meaning were also more likely to discount objective evidence that that meaning was not correct. In experiment 2, we explored whether confidence in a word not only influences learning for that particular word, but also influences the learning of other words. In particular, we show that confidence in the meaning of a word influences how likely learners are to use that word to make a mutual exclusivity inference (Markman & Wachtel, 1988), where a second new word is assumed to differ in meaning from that first learned word. Finally, experiment 3 is a pre-registered replication of the main findings of experiment 1 and experiment 2.

1 Experiment 1

As outlined in the introduction, subjective judgments of confidence can track the objective uncertainty that is present in the world. However, subjective confidence can also deviate from the objective uncertainty of the world. For instance, if people are highly confident in their decisions or knowledge, this can lead them to selectively choose information that is consistent with their beliefs while disregarding disconfirming information, a phenomenon known as confirmation bias (Nickerson, 1998). Confirmation biases have been demonstrated across a range of real-world scenarios, from the formation of political attitudes (Kaplan, Gimbel, & Harris, 2016; Lord, Ross, & Lepper, 1979), to low-level perceptual tasks (Rollwage et al., 2020; Talluri, Urai, Tsetsos, Usher, & Donner, 2018), suggesting that such a bias might also be active during word learning.

In experiment 1, we used confirmation bias as an assay, to test whether subjective confidence in having acquired the correct meaning of a word influences referent selection during a cross-situational statistical learning task. Participants completed a series of word learning trials in which we held constant the overall cross-situational learning statistics (the objective uncertainty surrounding a word meaning), but varied the order in which information was provided to induce a confirmation bias. Concretely, for each to-be-learnt word, participants were exposed to a series of 8 trials where a novel word was paired with four possible referents. While the correct referent (the *target*) appeared in all 8 trials, a *competitor* object appeared in 7 trials. To increase the likelihood that participants held an incorrect belief about the word meaning (i.e., choose the competitor object), we dynamically allocated the target and competitor based on the participant's response in the first learning trial: the competitor object was defined as the object selected on the first trial of the series such that participants always started with an incorrect belief about the word meaning. To induce a confirmation bias, we then manipulated the position of the single informative trial, i.e., the only trial that featured the target object but not the competitor: the informative trial either appeared as the second trial in the series (the early informative condition) or as the seventh (i.e. penultimate) trial (late informative condition; see Figure 1). Since participants started by selecting the competitor object, we hypothesized that, in the late informative condition, they would continue to select it when presented again (following previous work Dautriche & Chemla, 2014; K. Smith et al., 2011; Trueswell et al., 2013; Yurovsky & Frank, 2015), building up their confidence in this incorrect association, until they hit the informative trial. Accordingly, in the late informative condition, a learner who solely uses word-referent co-occurrence statistics would change their mind (i.e, switching to consistently choose the target object) upon realising their initial mistake, whereas a learner who displays a confirmation bias would maintain their incorrect belief despite the disconfirming evidence, and so return to selecting the competitor when it became available again. In contrast, in the early informative condition, no such confirmation bias is expected as the informative trial (the second trial in the series) arrives too early for participants to have built confidence in the word-competitor association.

We tested two signatures of an effect of confidence in referent selection in this task. First, we expected to find an order effect: participants should be more likely to display a confirmation bias when the informative trial appeared late, rather than early, in the series of trials. This is because in the late informative condition, they would have ample chance to confirm their incorrect belief before hitting the informative trial while in the early informative condition, the informative trial would appear too early for participants to hold high confidence in the word-competitor association. We thus expected that, at the end of the trial series, participants would be less likely to choose the target object in the late informative condition than in the early informative condition. Second, we expected to find the key signature of a confirmation bias: participants who hold high confidence in the word-competitor mapping would be less likely to use the informative trial to update their belief than participants with lower confidence scores.

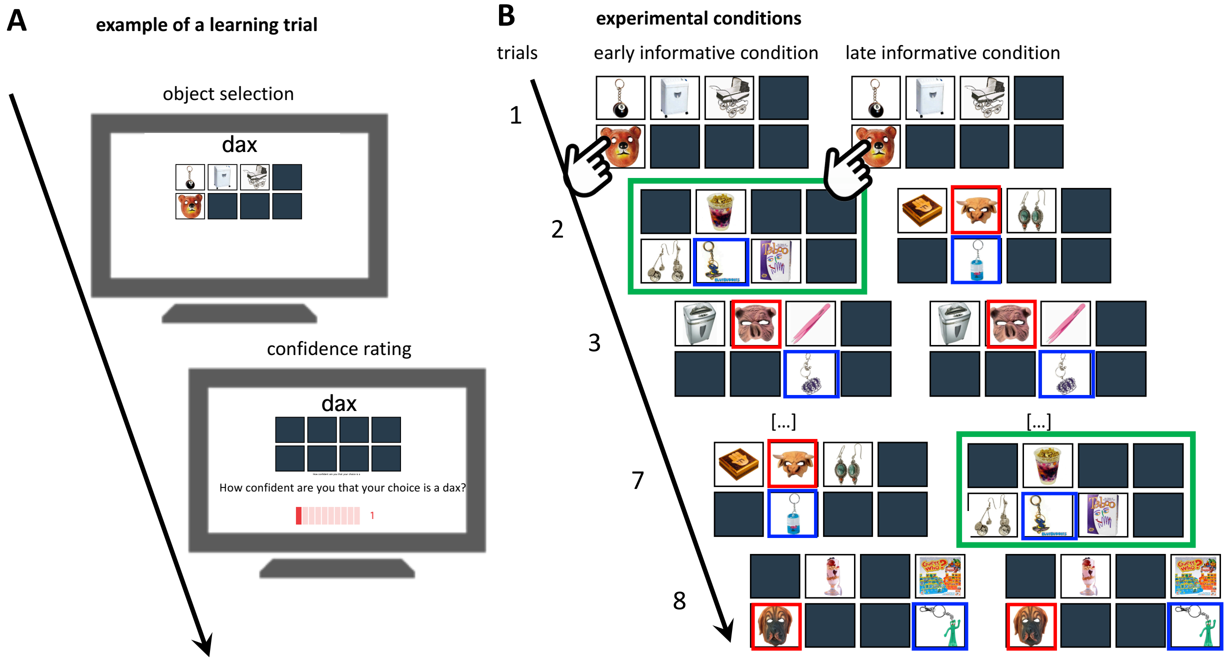
A
example of a learning trial
object selection
dax
confidence rating
dax
How confident are you that your choice is a dax?
1
B
experimental conditions
trials
early informative condition
late informative condition
1
2
3
[...]
[...]
7
8

Figure 1. A. Example of a learning trial. In each trial, 4 objects are displayed as possible meanings for the to-be-learnt word (here, "dax"). Participants first selected the object they believe best represents the meaning of the word. Once they responded, the pictures disappeared and they were asked to provide their confidence rating in their choice ("How confident are you that your choice is a dax?") on a 10 point scale displayed as a horizontal bar. **B. Experimental conditions.** Participants learnt words across a series of 8 trials. The target object (in blue) appeared in all 8 trials. A competitor object (in red) appeared 7 times. In the early informative condition, the informative trial (where the target appears but not the competitor; in green) appeared on the second trial in the series and for the late informative condition, it appeared on the seventh trial (late informative condition). The competitor and target objects were assigned dynamically during the experiments such that the object participants selected as the referent of the word in the first trial was the competitor and another unselected object appearing on the same trial became the target.

1.1 Method

The data and the script for their analysis are available here: https://osf.io/upndk/.

Participants. 49 adults were recruited through Amazon Mechanical Turk (all residing in the USA, all self-identified native speakers of English as per answers given on a questionnaire at the end of the experiment and all with a minimum of 50% approved HITs on AMT). Data collection proceeded in batches and stopped when at least 40 participants could be included in the final analysis; this sample size followed previous cross-situational learning studies (Dautriche & Chemla, 2014; K. Smith et al., 2011; Yurovsky, Yu, & Smith, 2013, all having 40 to 50 participants). Three participants were excluded from the analysis because: they encountered technical issues (n = 2) or they reported using a pen to track their referent selection at each trial (n = 1). See exclusion criteria below for further details, including details of trial-by-trial exclusions. The final sample consisted of 46 participants (20 females, age 20 to 59 years, mean age 34 years).

Procedure. Participants were tested online. They were instructed that they were to learn words by associating them with images displayed on the screen. Prior to test, participants were given a screenshot of a learning instance involving a word and a set of pictures that were not used at test. No information was given about the number of to-be-learned words or the number of trials per word. For each trial, participants were asked to click on the image they believed best represents the meaning of the word. Once they responded, the pictures disappeared and they were asked to rate their confidence in their choice on a 10 point scale displayed as a horizontal bar (see Figure 1A). Once they clicked on a point in the scale, the test continued with the next trial. We recorded participants' choice and confidence at each trial as well as their response times. Participants had as much time as they wanted to give their responses. A final questionnaire asked for participants' gender, age, native language and whether they used a pen during the experiment. The experiment lasted around 10 min, and participants were paid $1.50.

Design and stimuli. Participants learnt a total of 6 words across a series of learning trials in a cross-situational learning design (see, e.g., Dautriche & Chemla, 2014). The word labels were six phonotactically legal English non-words: "blicket", "tupa", "dax", "moop", "zud", "smick". There were 8 learning trials per word, resulting in a total of 48 trials for the whole experiment. The words were learnt serially, such that the 8 learning trials for a given word were presented one after another, followed by the 8 learning trials for the next to-be-learnt word and so on (e.g., 8 trials for "blicket", followed by 8 trials for "tupa", etc). On each learning trial participants saw the word, e.g., "blicket", above a 2x4-cell grid in which 4 possible object referents were displayed at a random spatial location. For all words, the 4 object pictures presented in a learning trial were selected pseudo-randomly from a bank of images (`https://bradylab.ucsd.edu/stimuli.html`, Konkle, Brady, Alvarez, & Oliva, 2010). Objects that appeared in several learning trials for a given word (such as the target object) appeared with a different picture in each trial. For instance, a word whose meaning could be glossed as *dog* would be pictured as a German Shepherd on the first trial, as a Dalmatian on the second trial, etc.

Participants had to learn 2 critical ambiguous words and 4 filler words, which were included in order to minimize the risk that participants would become aware of the structure of the critical ambiguous word trial series over the course of the experiment. The filler and critical words series were constructed as follow:

- For two of the filler words, the target appeared on all trials and all distractors appeared exactly once. For the other two filler words, all objects appeared 4 times across the 8 trials, such that they were all equally likely to be the target.
- For the critical ambiguous words, the target object appeared in all 8 trials. A competitor object appeared 7 times. For one word, the informative trial (where the target appears but not the competitor) appeared on the second trial in the series (early informative condition) and for the other word, it appeared on the seventh trial (late informative condition). The competitor and target objects were assigned dynamically during the experiments such that the object participants selected as the referent of the word in the first trial was the competitor and another unselected object appearing on the same trial became the target.

There was no overlap between the set of objects used for different words, ensuring that participants could not infer the meaning of words through mutual exclusivity (Markman & Wachtel, 1988). The one-to-one pairing between words and object referent types (target, competitor and distractors) was fully randomized and differed for each participant.

The experiment always started with one of each type of fillers. Filler words were used to ensure that participants understood the procedure while being exposed to varying degrees of difficulty.

Criteria for exclusion. We excluded trials for which no response or target object was saved (27 trials excluded; 1.2% of the total number of trials) or for which responses times were implausibly fast (less than 2000ms for selecting a referent and giving a confidence judgement, no trial excluded; mean response time = 8856ms, SE=313ms). We excluded a series of learning trials (corresponding to a given word) when: participants provided a confidence level greater or equal to 5 out of 10 on the first trial (3 trial series excluded) and when participants had less than 4 usable trials in the series (i.e., were missing half of the trials due to individual trials being excluded; 2 series of trials/words excluded). Participants were then excluded if they used a pen during the experiment (given their response in the final questionnaire; n = 1), had technical issues they reported (e.g., pictures not displayed; n = 2) or had less than half of the data for ambiguous words available after trial- and word-based exclusion (n = 0).

Data analysis. We performed mixed model analyses with the `lme4` package (v.1.1-21) in R (Bates, Mächler, Bolker, & Walker, 2014). We used the maximal random effect structure as supported by the data. P values for main fixed effects are based on likelihood ratio tests (Dobson & Barnett, 2008), simple effects are reported from the summary table of the model. Models are reported in the Appendix following the format recommended by Meteyard and Davies (2020).

1.2 Results

Preliminary analysis: design validation. Figure 2A presents the proportion of competitor responses across the series of trials for each condition (late informative, early informative; results for the filler words can be found in the online analysis script). Our design rests on the assumption that participants in the late informative condition would continue selecting the competitor on trials 1–6 before reaching the informative trial 7, while participants in the early informative condition would switch from the competitor to the target object after seeing the informative trial 2, and then select the target consistently for the remainder of the experiment (trials 3 to 8). We restricted our analysis to trials 3 to 6 where both the competitor and target objects were present and available for selection in both conditions. A mixed-effects model with random intercepts for each participant and fixed effects for Condition and TrialNumber showed a significant main effect of Condition on competitor responses ($\chi^2(1)$ = 107.59; p < .001; Cohen's d = 1.31; see Appendix Table A.1). Our manipulation worked as expected: in the late informative condition, participants were more likely to select the competitor object on trials 3–6, hence holding an incorrect belief about the word-referent mapping, than in the early informative condition.

As can be seen in Figure 2B, there was no difference in participants' confidence across conditions from trial 1 to 6 despite the word meaning being objectively ambiguous in the late informative condition (the target and the distractor both appeared equally often with the word) but unambiguous in the early informative condition (the target co-occured 6 times with the word while the competitor co-occured 5 times) (no main effect of Condition, $\chi^2(1)$ = 1.74; p = .19, and no interaction between TrialNumber and Condition, $\chi^2(1)$ = 0.18; p = .67, in a mixed-effect model with random intercepts for each participant; see Appendix Table A.2). This suggests that participants' confidence scores were higher than warranted by the objective uncertainty of the world in the late informative condition, which would be consistent with a confirmation bias.

In sum, our manipulation was successful in producing the right situation for the observation of a confirmation bias where participants held an incorrect belief with strong confidence.

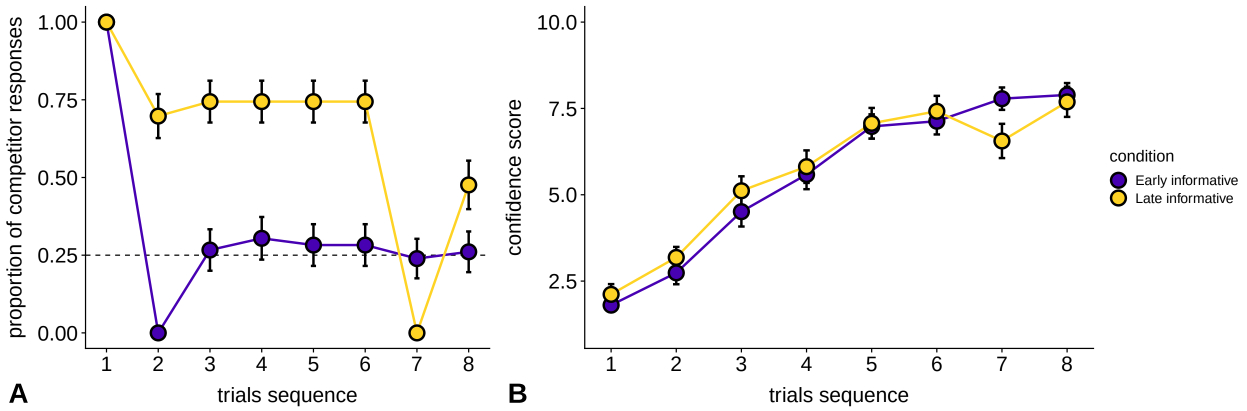

Figure 2. Competitor selection rate and confidence across trials. A. Proportion of competitor responses as a function of trial position in the learning series for each condition (Early informative vs. Late informative). Participants' first response (on trial 1) was defined as the competitor object (proportion of competitor responses = 1). The competitor object appeared then 7 times across the series of 8 learning trials for that word. The target appeared in all trials. The informative trial, i.e., the only trial where the competitor object did not appear, was either trial 2 (early informative condition) or trial 7 (late informative condition) (both proportion of competitor responses = 0). Participants selected the competitor object at a higher rate in the late informative condition compared to the early informative condition, thus holding an incorrect belief about the word-referent mapping. The dashed line represents the chance level (0.25). **B.** Average confidence scores as a function of trial position in the learning series for each condition. Error bars represent the standard error of the mean. Overall confidence scores were fairly similar across conditions throughout the trial series but in trial number 7 in the late informative condition (corresponding to the informative trial), participant's confidence dropped as the competitor object (their most likely selection in the previous trial) was not available for selection.

Analysis 1: Order effect. Figure 3A shows the average proportion of target and competitor selections by condition (early informative vs. late informative) in the final (8th) trial. The selection rate of the two distractor objects (i.e. the objects presented on trial 8 which were neither the target or the competitor, not plotted) was low and comparable across conditions ($M_{early} = 0.02, SE_{early} = 0.02$, $95\%CI = 0.04$; $M_{late} = 0.02, SE_{late} = 0.02$, $95\%CI = 0.05$; $\chi^2(1) = 0.004$; $p = .95$), thus we treated participants' target and competitor responses as complementary and analyzed participants' target response as a function of Condition with random intercept for each participant in our mixed-model analysis. We found a significant effect of Condition, i.e., position of the informative trial, on participants' responses in the final trial: in the early informative condition participants selected the target object ($M = 0.72, SE = 0.07, 95\%CI = 0.13$) significantly more than in the late informative condition ($M = 0.50, SE = 0.08, 95\%CI = 0.16$; $\chi^2(1) = 4.61$; $p = .03$; Cohen's $d = 0.67$; see Appendix Table A.3).

As can be seen in Figure 3B, the order in which trials were seen also affected participants' confidence in their final response, depending on whether they chose the target or the competitor. A mixed-effect model with random intercept for each participant showed a significant interaction of chosen Object Type (target vs. competitor) and Condition (Early vs. Late) on confidence scores ($\chi^2(1) = 8.42$; $p = .004$; Cohen's $d = 0.70$; no other effect was significant; see Appendix Table A.4): In the late informative condition, participants who (objectively, incorrectly) chose the competitor object were more confident in their response than those who chose the target object ($M_{competitor} = 8.65, SE_{competitor} = 0.45, 95\%CI = 0.95$; $M_{target} = 6.71, SE_{target} = 0.70, 95\%CI = 1.47$; $\beta = 1.86, t = 2.59, p = .01$). By contrast, in the early informative condition, participants gave higher confidence scores after selecting the target object than after selecting the competitor object, although this difference was not significant ($M_{competitor} = 7.08, SE_{competitor} = 0.69, 95\%CI = 1.52$; $M_{target} = 8.36, SE_{target} = 0.35, 95\%CI = 0.72$; $\beta = 1.16, t = 1.49, p = .14$).

Such order effects have also been previously reported elsewhere (Medina et al., 2011; Thaker, Tenenbaum, & Gershman, 2017), and have been accounted for by participants' past selection history consistent with an hypothesis-testing account (K. Smith et al., 2011; Stevens, Gleitman, Trueswell, & Yang, 2017; Thaker et al., 2017; Trueswell et al., 2013; Yurovsky & Frank, 2015). Here, however, we suggest that this is a result of participants being more likely to display a confirmation bias in the late informative condition when they hold high confidence in the competitor object. The next analysis provides a more direct test for this hypothesis.

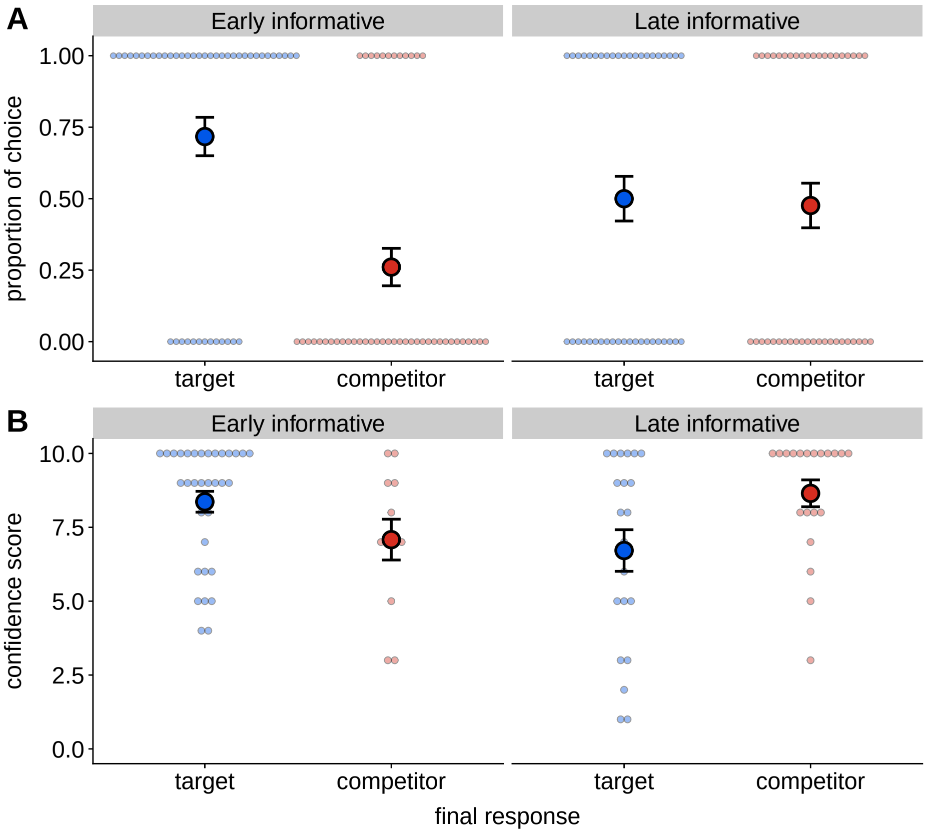

Figure 3. Analysis 1: Order effect. Selection rate **(A)** and associated confidence scores **(B)** of the target and competitor objects in the final trial. Dots represent participants' individual response and error bars represent the standard error of the mean.

Analysis 2: Confirmation bias. We next tested for specific evidence of a confirmation bias effect in word learning: When people are highly confident in the (incorrect) word-competitor mapping, are they more likely to discount disconfirming information?

We focused on the late informative condition and selected all responses where participants chose the competitor in trial 6 (prior to the informative trial) and chose the target during the informative trial (n = 30, of whom 18 switched back to the competitor in the final trial). Figure 4 shows the average confidence score in trial 6 after selecting the competitor as a function of the final response (target vs. competitor). Participants' who gave higher confidence ratings on trial 6 were then more likely to revert back from the target to the competitor on trial 8, despite having received evidence against the word-competitor mapping on trial 7. This was reflected by a main effect of Confidence on participants' final selection ($\chi^2(1)$ = 10.97; p < 0.001; Cohen's d = 1.38; see

Appendix Table A.5). This effect was not explained by the number of times participants selected the competitor across trials 1 to 6 (the effect of confidence when controlling for this was still significant; $\chi^2(1)$ = 12.09; p < 0.001).

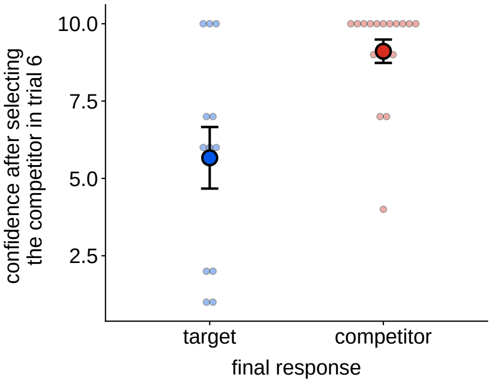

Figure 4. Analysis 2: Confirmation bias. Average confidence score in the word-competitor mapping prior to the informative trial as a function of the final response (target vs. competitor) in the late disambiguation condition. Dots represent participants' individual response and error bars represent the standard error of the mean.

1.3 Summary of Experiment 1

Experiment 1 provides evidence that the process of cross-situational word learning is influenced by the learner's subjective confidence in an acquired meaning. Specifically, our results show an effect of input order: Participants were less likely to reach the correct word-object mapping when they received an informative trial late in the learning process as compared to when they received it early, despite the word-object co-occurrence statistics being preserved overall. Moreover, when participants were more confident in an incorrect word-object mapping, then they were also more likely to ignore an informative trial that disconfirmed the mapping. These findings are consistent with the existence of a confirmation bias during word learning: participants who had the chance to confirm an incorrect word-object mapping multiple times (as in the late informative condition) were more likely to experience a boost in confidence for that mapping and thus more likely to persist in their belief in that mapping even after receiving objective evidence that they were incorrect.

2 Experiment 2

Experiment 1 showed that subjective confidence influences information processing during the learning of individual words. However, it has long been known that learners do not learn words individually, but as part of a lexical system (e.g., O'Hanlon & Roberson, 2006; Tillman & Barner, 2015; Waxman & Klibanoff, 2000). Most famously, a large literature attests to the existence of *mutual exclusivity* effects (Markman & Wachtel, 1988), according to which a new word cannot share the same meaning as an existing word. In a task such as ours, mutual exclusivity can be used to acquire a word's meaning in a single learning instance, if all the distractor objects are already associated with a label. Indeed, both child and adult learners will assume that a novel label refers to a novel object, rather then to a familiar object for which they already have a label (e.g., Diesendruck & Markson, 2001; Golinkoff, Hirsh-Pasek, Bailey, & Wenger, 1992; Graham, Poulin-Dubois, & Baker, 1998; Halberda, 2003; Xu, Cote, & Baker, 2005; Yurovsky & Yu, 2008). Thus, associating a meaning with a word will have consequences beyond that single pairing, as it can influence subsequent inferences made about other words in the lexicon and thus accelerate the word learning process.

An important question is how such a principle of mutual exclusivity works when word knowledge is imperfect, as ought to be the case when words are learnt across multiple ambiguous exposures in a cross-situational fashion. Many computational models of cross-situational word learning capture mutual exclusivity effects, i.e., they can rule out a familiar meaning such as cat as the referent of a novel word "blicket", without building an explicit assumption of mutual exclusivity, simply by using co-occurrence regularities, i.e., the probability of using "blicket" in the presence of a cat is lower than the probability of using "cat" (Fazly et al., 2010; S. Frank et al., 2009). These models therefore suggest that mutual exclusivity effects may appear as a consequence of using cross-situational learning statistics. However, this could result in learners applying mutual exclusivity even if they had a very incomplete knowledge of the word "cat", as even after very few exposures the probability of co-occurence of "cat" and its correct referent would still be higher than of "cat" with a novel object. Recent experimental evidence suggests that this is not the case: toddlers are more likely to use a familiar word to make a mutual exclusivity inference if they had more experience with that word (Lewis, Cristiano, Lake, Kwan, & Frank, 2020); most tellingly, a single exposure to a word meaning does not seem to trigger mutual exclusivity inferences based on that word. Thus the degree of familiarity with a word seems to influence the extent to which one applies mutual exclusivity. Following this, we hypothesized that learners' *subjective* word knowledge should influence the application of mutual exclusivity over and beyond participants' experience with that word.

In experiment 2, we tested whether learners' subjective confidence in the word-meaning mapping influences mutual exclusivity inferences based on that word while controlling for the objective knowledge they accumulated about that word. Similarly to experiment 1, participants learnt words across a series of learning trials in a cross-situational statistical learning task. To induce different confidence levels, we manipulated the difficulty of the series of learning trials: some words were easy to learn because no distractor co-occurred more than once with the word, whereas other words were harder as all distractors co-occured multiple times with the word. At the end of the trial series for a given word, e.g., "dax", participants were presented with a first mutual exclusivity trial where a novel word, "roplixoo", was presented alongside 1) the object they most recently selected as the likely meaning for "dax" and 2) another object that did not have a label. We tested whether participants' confidence in the most-recently selected meaning of "dax" would predict if they generate mutual exclusivity inferences in this trial, i.e. selecting the unlabeled object in response to the novel word "roplixoo". Participants were then presented with a second mutual exclusivity trial, where the novel target seen on the first mutual exclusivity trial was presented together with another unlabelled object and another novel word, e.g., "bosa". We tested whether participants' confidence during the first mutual exclusivity trial, i.e., their confidence that they had inferred the meaning of "roplixoo", predicts a mutual exclusivity inference in this trial, even in the absence of direct evidence for the meaning of the word "roplixoo" (see Figure 5).

2.1 Method

The data and the script for their analysis are available here: https://osf.io/upndk/.

Participants. 45 adults were recruited through Amazon Mechanical Turk. Data collection proceeded in batches and stopped when at least 40 participants could be included in the final analysis, as per experiment 1. One participant was excluded from the analysis because more than

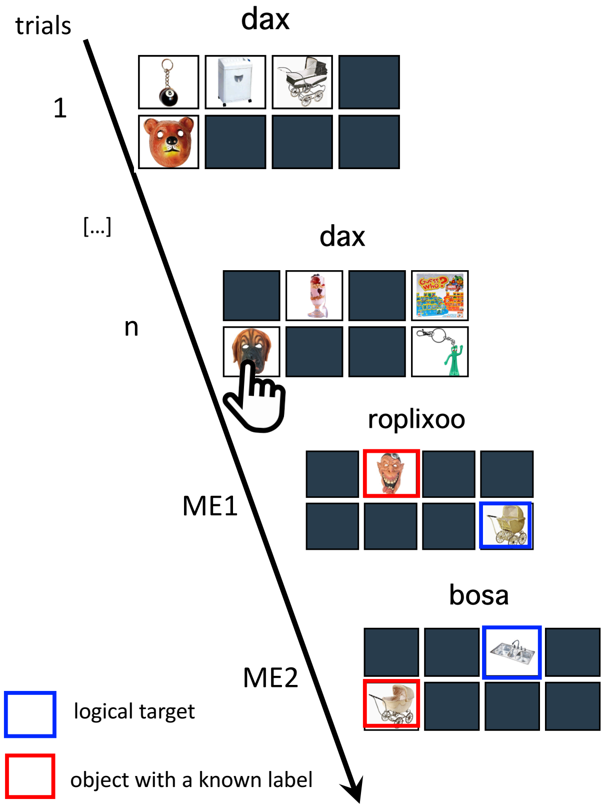

Figure 5. Structure of the trials series for experiment 2 and 3. Participants learnt a first word w, e.g., "dax", across a series of n trials (n = 6 in experiment 2 and n = 8 in experiment 3). At the end of the trial series, participants were presented with a first mutual exclusivity trial (ME1) where a novel word w_{ME1} is presented along with the object that they selected in the immediately preceding trial as the meaning of dax (w last referent) and an object that did not have any label, thus the logical target for w_{ME1}. Participants were then presented with a second mutual exclusivity trial (ME2), where the w_{ME1} logical target was presented along with another unlabeled object, hence the logical target for w_{ME2}.

half of their data was not exploitable after trial exclusion. See details in the exclusion criteria section. The final sample consisted of 44 participants (20 females; age 20 to 68 years, mean age 36 years, all residing in the USA, 43 self-identified native speakers of English as per answers given on a questionnaire at the end of the experiment and all with a minimum of 50% approved HITs on AMT)

Procedure. Same as experiment 1.

Design and Stimuli. Participants learnt a total of 10 words, each word presented across a series of 6 learning trials followed by two mutual exclusivity trials (see Figure 5). Participants had to learn 3 easy words, 3 hard words and 4 filler words; the experiment always started with a filler word. For all words the target object appeared in all learning trials for that word.

- For the easy and the filler words, all distractors appeared exactly one time.
- For the hard words, all distractors appeared 3 times across the 6 learning trials.

The 6 learning trials for each word w were then followed by two mutual exclusivity trials using two novel words, w_{ME1} and w_{ME2}. The timing and presentation of the mutual exclusivity trials were similar to the learning trials, except that only two possible object referents were presented alongside the novel word. For the easy and hard words, the two mutual exclusivity trials were composed as follows:

- In the first mutual exclusivity trial (ME1), one object was the object selected in the immediately preceding trial (trial 6; the participant's last hypothesis for w) and the other object was a distractor participants saw but never selected during the 6 learning trials for w; for a learner applying mutual exclusivity who believes they have correctly learnt the meaning of w, this referent is therefore the logical target for w_{ME1}.
- In the second mutual exclusivity trial (ME2): one object was the target of w_{ME1} and, similarly to the first mutual exclusivity trial, the other object was a distractor participants saw but never selected during the learning trials for w (therefore the logical target for w_{ME2}).

For filler words, both ME trials featured two distractor objects which had been seen but not selected during the learning trials for that filler word. The role of the filler words was to display some variability in the ME trials such that the immediately preceding selection was not always featured in these trials. In ME trials, each word was displayed using a different color in order to minimize the risk that participants ignore the word when giving their response. The rest was similar to experiment 1.

Criteria for exclusion. The criteria for exclusion were the same as in experiment 1. We excluded trials for which no response or target object was saved (n = 11 trials) or for which responses times were implausibly fast (less than 2000ms, n = 29; mean response time = 5970ms, SE=141ms). We excluded a series of learning trials (corresponding to a given word) when: participants started with a confidence level greater or equal to 5 (n = 20 of trials/words excluded) and when participants had less than 3 usable trials in the series (half of the trials; n = 1). Participants were then excluded if they used a pen during the experiment (given their response in the final questionnaire; n = 0), had technical issues they reported (e.g., pictures not displayed; n = 0) or had less than half data on easy and hard words available after trials and word exclusion (n = 1).

Data analysis. Same as experiment 1.

2.2 Results

Preliminary analysis: Design validation. We first inspected participants' performance and confidence at the end of the easy and hard 6-trial learning series, prior to the ME tests. Participants displayed better performance in choosing the correct referent for easy (M = 0.93, SE = 0.02, 95%CI = 0.06) compared to hard words (M = 0.78, SE = 0.04, 95%CI = 0.09), as reflected by a main effect of trial difficulty on target selection proportion ($\chi^2(1)$ = 14.53; p < .001; see Appendix Table A.6). This was also reflected on confidence scores, with participants being on average more confident for easy words (M = 8.30, SE = 0.35, 95%CI = 0.70) compared to hard words (M = 7.22, SE = 0.31, 95%CI = 0.63; $\chi^2(1)$ = 12.73; p < .001; see Appendix Table A.7). Thus our manipulation of word learning difficulty worked as expected in producing variable performance and confidence levels prior to presenting participants with mutual exclusivity trials.

Analysis 3: ME1. Participants' application of mutual exclusivity was not modulated by w's learning difficulty: Participants overwhelmingly selected the object they did not already associate a label to (M_{easy} = 0.85, SE_{easy} = 0.04, 95%CI = 0.08; M_{hard} = 0.89, SE_{hard} = 0.03, 95%CI = 0.06; no main effect of difficulty on ME accuracy $\chi^2(1)$ = 1.02, p = .31). Thus, we analyzed all ME1 trials altogether irrespective of w's learning difficulty. Figure 6A shows the average confidence score in the last referent selected for w in the immediately preceding trial as a function of the response in the ME1 trial (w_{ME1} target vs. w last referent). Participants who were more confident in having learnt w were more likely to choose the target than their w last referent, as reflected by a main effect of Confidence on participants' selection in a mixed-effect model with Participant as a random effect ($\chi^2(1)$ = 18.63; p < .001; Cohen's d = 0.70; see Appendix Table A.8). Critically, this effect persisted when controlling for the number of times the w referent was selected throughout the 6 learning instances ($\chi^2(1) = 4.10$; $p = .04$; Cohen's $d = 0.46$).

Analysis 4: ME2. We analyzed all ME2 trials following a correct ME1 trial (86% of all ME2 trials). Participants successfully applied mutual exclusivity on ME2 trials despite having not received direct evidence for the w_{ME1}-target mapping: They selected the object that did not have a label associated with it, as opposed to the w_{ME1} target, at a rate greater than chance ($M = 0.78$, $SE = 0.03$, $95\%CI = 0.06$; $\beta = 1.27, z = 6.18, p < .001$). Figure 6B shows the average confidence score in the w_{ME1}-target mapping in the ME1 trial as a function of the response in the ME2 trial (w_{ME2} target vs. w_{ME1} target). Participants were more likely to choose the w_{ME2} target when they had higher confidence in the w_{ME1}- target link ($\chi^2(1) = 5.78$; $p = .02$; Cohen's $d = 0.33$; see Appendix Table A.9). Importantly, in this analysis, participants had all made the same choice on the previous relevant trial (w_{ME1}), indicating that confidence makes a unique contribution to mutual exclusivity inferences.

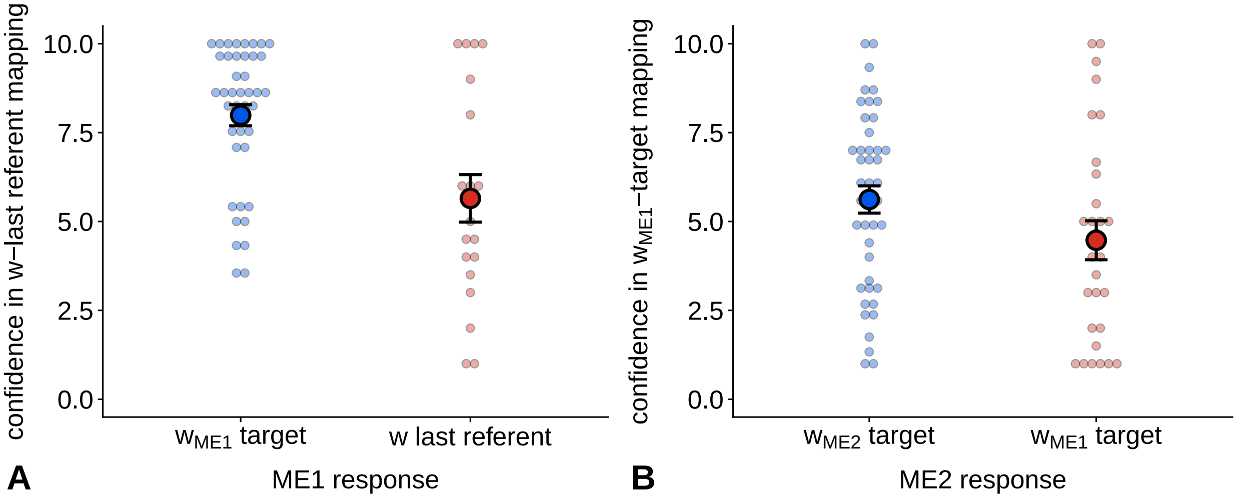

Figure 6. A. Analysis 3: ME1. Average confidence score in the w-last selected referent mapping as a function of the response in the ME1 trial (w_{ME1} target vs. w last referent).**B. Analysis 4: ME2.** Average confidence score in the w_{ME1}-target mapping as a function of the response in the ME2 trial (w_{ME2} target vs. w_{ME1} target). Dots represent participants' individual response and error bars represent the standard error of the mean.

2.3 Summary of Experiment 2

Experiment 2 provides evidence that confidence in a word meaning influences how likely learners are to use this word to make mutual exclusivity inferences. Participants were more likely to apply mutual exclusivity, i.e. map a new label to an object that does not have a label, when they were confident that they knew the label of the alternative possible referent. Critically, subjective confidence in a word's meaning influenced mutual exclusivity inferences beyond participants' objective experience with the word-meaning mapping (i.e., confidence had an effect above-and-beyond the number of times the participants selected an object as the likely meaning of the word), and it also had an effect in the absence of direct evidence for the word-meaning mapping (as in the second mutual exclusivity trial). In sum, our results suggest that subjective confidence in a word-meaning mapping not only influences the information selection process for that word, as shown in experiment 1, but also modulates inferential processes based on that word, affecting learning elsewhere in the lexicon.

3 Experiment 3

Experiment 3 aimed to replicate the main findings of experiments 1 and 2 with a larger sample size, giving us better opportunity to estimate the relevant effect sizes, and minimise the possibility of false positives or negatives. The experimental protocols and all analyses were preregistered. As in experiment 1, we tested whether word learners are sensitive to information order during word learning (Analysis 1), and display a confirmation bias (Analysis 2), two signatures of an effect of subjective confidence in referent selection. Following experiment 2, we also tested whether subjective confidence in a word meaning affects subsequent mutual exclusivity-based inferences (Analyses 3 and 4).

3.1 Method

The pre-registration, the data and the script for their analysis are available here: https://osf.io/upndk/.

Participants. 144 adults were recruited through Amazon Mechanical Turk. 69 participants were excluded from the analysis because: they used a pen to track their referent selection at each trial ($n = 50$), they failed to provide 80% of correct answers on control trials ($n = 10$) or more than half of their data was not exploitable after trial exclusion ($n = 9$) (all pre-registered criteria). It is to be noted that this experiment was run during the lockdown imposed by COVID (summer 2020), as a result the pool of participants may be critically different from the ones of experiments 1 and 2, which may explain the high rejection rate (although see Moss, Rosenzweig, Robinson, & Litman, 2020). The final sample consisted of 75 participants (31 females; $M = 38, min = 22, max = 70$ years of age, all native speakers of English). The required number of participants was estimated by using experiment 1's data. Using the R package `pwr`, our power analysis based on the order effect (difference in final trial accuracy between the late and the early disambiguation condition) suggested that we should test 70 participants to have a power of 80% at the 0.01 alpha level assuming a large effect side $d = .8$.

Procedure. Same as experiment 1.

Design and Stimuli. The design combined the manipulations of experiment 1 and experiment 2. Participants learnt a total of 10 words, each presented in a series of 8 learning trials followed by 2 ME trials (see Figure 5). Participants had to learn 2 ambiguous words with an early informative trial, and 2 with a late informative trial (as in Experiment 1), and 3 easy words and 3 hard words (as in Experiment 2). For all words the target object appeared in all learning trials for that word.

- For the ambiguous words, see method section of Experiment 1.
- For the easy words, all distractors appeared exactly one time.
- For the hard words, all the objects appeared 3 times.

The 8 learning trials for word w were then followed by two mutual exclusivity trials, as in experiment 2. During the first mutual exclusivity trial (ME1), participants were presented with a novel word w_{ME1} and two object pictures on the screen:

- For the easy and hard words: the object the participant clicked on in trial 8 (i.e. their last hypothesis for w) and a distractor they saw but never clicked during the preceding 8 learning trials (therefore the logical target for w_{ME1} for a learner applying mutual exclusivity).
- For the ambiguous words: the competitor and a distractor they saw but never clicked on during the preceding 8 learning trials.

During the second mutual exclusivity trial (ME2), again participants are presented with a novel word, w_{ME2}, and two objects on the screen: the object that participants clicked on in the preceding ME1 trial and a distractor they never clicked on but saw during the 8 learning trials of w.

In addition we distributed 10 catch trials across the whole experiment, to confirm that participants were attending to the task. These catch trials were similar to the learning trials: they featured 1 familiar word ("apple", "ball", "coin", "dog", "hat", "key", "leaf", "pizza", "plane", "watch") and 4 object pictures. There was always exactly one catch trial added to the 8 learning series for each word w. These catch trials were added after informal discussions with peers who warned us about the poor quality of AMT's data during the summer of 2020.

For the sake of clarity we summarize below the differences between experiment 3 and experiments 1 and 2:

- The presence of catch trials (and the exclusion criteria based on the responses to these trials, see below).
- Participants learnt 4 ambiguous words (instead of 2 in experiment 1).
- For hard words, distractors appear 4 times (instead of 3 times as in experiment 2) as there are now 8 learning instances (instead of 6 in experiment 2).
- In ME2, the object participants selected during ME1 appeared in ME2 (whereas in experiment 2, the target of ME1 appeared in ME2).

Criteria for exclusion. The criteria for exclusion were the same as in experiment 1 and 2. We excluded trials for which no response or target object was saved (n = 59 trials) or for which responses times were implausibly fast (less than 2000ms, n = 53; mean response time = 6260ms, SE=86ms).

We excluded a series of learning trials (corresponding to a given word) when: participants started with a confidence level greater or equal to 5 and when participants had less than 4 usable trials in the series (n = 146). Participants were then excluded if they used a pen during the experiment (n = 50), had less than half data on easy and hard words available after trials and word exclusion (n = 9). In addition, we added an exclusion criteria on the control words: we rejected participants who failed to answer correctly on 2 or more control trials (out of 10) and/or who gave a confidence rating of less than 8 for 2 or more of these control trials (n = 10).

Data analysis. Same as experiment 1. Note that all pre-registered mixed models analyses only included a random intercept for participants. Yet because the model with a random slope Condition per participants converged (contrary to Experiment 1), we included a random slope for Condition per participant in our models (see details of the models in the Appendix).

3.2 Results

Experiment 3 confirmed and replicated the major findings from Experiments 1 and 2.

Analysis 1 (replication): Order effect. Figure 7A shows the mean proportion of target and competitor object selections for each ambiguous word condition (early informative vs. late informative) in the final trial. Because the selection rate of the distractor objects (not represented in the figure) was small and comparable across conditions (M_{early} = 0.08, SE_{early} = 0.03, 95%CI = 0.06; M_{late} = 0.12, SE_{late} = 0.03, 95%CI = 0.06; $\chi^2(1)$ = 0.76; p = .38), we treated participants' target and competitor responses as complementary as in experiment 1 and analyzed participants' target response as a function of Condition with a random intercept and a random slope for Condition per participant in our mixed-model analysis (as participants now provided multiple trials per condition). Participants in the early informative condition selected the target object (M = 0.66; SE = 0.05, 95%CI = 0.10) significantly more than in the late informative condition (M = 0.50, SE = 0.08, 95%CI = 0.10; $\chi^2(1)$ = 6.55; p = .01, Cohen's d = 0.69 see Appendix Table A.10).

As can be seen in Figure 7B, the order in which trials were seen also affected participants' confidence in their final response. A mixed-effect model (with a random slope for Condition for each Participant) showed a main effect of Condition ($\chi^2(1)$ = 8.46; p = .004; Cohen's d = 0.60; see Appendix Table A.11) and a marginal interaction between ObjectType (target vs. competitor) and Condition (Early vs. Late) on confidence scores ($\chi^2(1)$ = 2.87; p = .09; Cohen's d = 0.24). In the early informative condition, participants who chose the target object were more confident in their response than those who chose the competitor object ($M_{target} = 7.77, SE_{target} = 0.29$, 95%$CI$ = 0.58; $M_{competitor} = 6, SE_{competitor} = 0.54$, 95%$CI$ = 1.10; $\beta = 1.69, t = 3.33, p = 0.001$). In contrast, in the late informative condition, participants' confidence score was similar for target and competitor responses ($M_{target} = 6.57, SE_{target} = 0.40$, 95%$CI$ = 0.80; $M_{competitor} = 6.27, SE_{competitor} = 0.44$, 95%$CI$ = 0.89; $\beta = 0.5, t = 0.97, p = 0.33$).

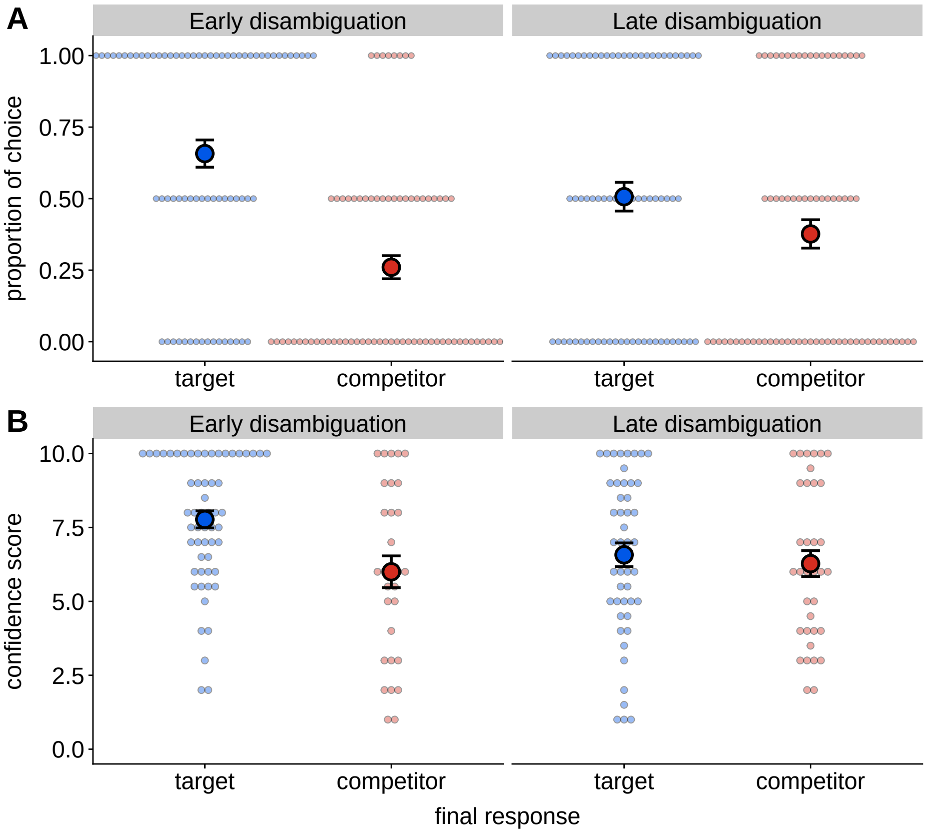

Figure 7. Analysis 1 (replication): Order effect. Selection rate **(A)** and associated confidence scores **(B)** of the target and competitor objects in the final trial. Dots represent participants' individual response and error bars represent the standard error of the mean.

Analysis 2 (replication): Confirmation bias. Figure 8 shows the average confidence score in trial 6 (late disambiguation condition) after selecting the competitor as a function of the final response (target vs. competitor). Participants who displayed higher confidence in the competitor response prior to the disambiguation trial were more likely to choose the competitor than the target in the final trial despite having received evidence against the word-competitor mapping. A mixed-effect model with a by-participant random intercept showed a significant main effect of Confidence on participants' final selection ($\chi^2(1)$ = 11.5; $p < .001$; Cohen's d = 0.71) that persisted even when adding the number of competitor selections across trials 1 to 6 as a predictor ($\chi^2(1)$ = 11.81; $p < .001$; Cohen's d = 0.71; see Appendix Table A.12).

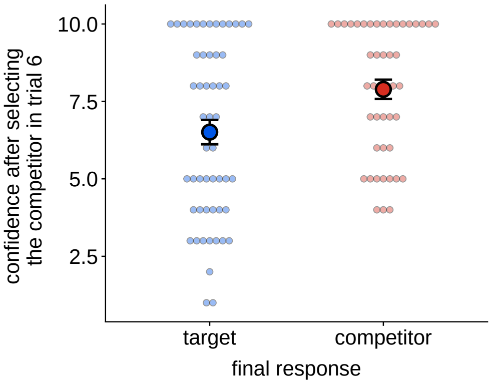

Figure 8. Analysis 2 (replication): Confirmation bias. Average confidence score in the word-competitor mapping prior to the informative trial as a function of the final response (target vs. competitor) in the late disambiguation condition. Dots represent participants' individual response and error bars represent the standard error of the mean.

Analysis 3 (replication): ME1. Our analysis here not only incorporated the easy and hard word trials (as in Experiment 2) but also the ambiguous word trials (which were not present in Experiment 2). For the ambiguous words, we selected all responses where participants chose the competitor object in trial 8, prior to ME1 trial, as this is the object that would be presented in the ME1 trial. Because participants' application of mutual exclusivity did not differ across word types ($M_{ambiguous} = 0.85, SE_{ambiguous} = 0.04$, 95%$CI$ = 0.09; $M_{easy} = 0.89, SE_{easy} = 0.03$, 95%$CI$ = 0.06; $M_{hard} = 0.84, SE_{hard} = 0.03$, 95%$CI$ = 0.06; $\chi^2(3)$ = 3.45, p = .33), we analyzed all ME1 trials altogether irrespective of w's type. Figure 9A shows the average confidence score in the last referent selected for w in the immediately preceding trial as a function of the response in the ME1 trial (w_{ME1} target vs. w last referent). Participants who were more confident in having learnt *w* were more likely to apply mutual exclusivity, as reflected by a main effect of Confidence on participants' selection in a mixed-effect model with Participant as a random effect and participant past selection history of the *w* referent (i.e., the number of times they selected it throughout the 6 learning instances) as a fixed effect ($\chi^2(1) = 17.23$; $p < .001$; Cohen's $d = 0.46$; see Appendix Table A.13).

Analysis 4 (replication): ME2. We analyzed all ME2 trials following a correct ME1 trial (80% of all ME2 trials). Participants successfully applied mutual exclusivity despite not having received direct evidence for the w_{ME1}-target mapping: They selected the object that did not have a label associated with it, as opposed to the w_{ME1} target, at a rate greater than chance ($M = 0.82$, $SE = 0.02$, $95\%CI = 0.03$; $\beta = 1.72, z = 9.7, p < .001$). Figure 9B shows the average confidence score in the w_{ME1}-target mapping in the ME1 trial as a function of the response in the ME2 trial (w_{ME2} target vs. w_{ME1} target). Participants were more likely to choose w_{ME2} target when they had higher confidence in the w_{ME1}- target link ($\chi^2(1) = 10$; $p = 0.001$; Cohen's $d = 0.34$; see Appendix Table A.14).

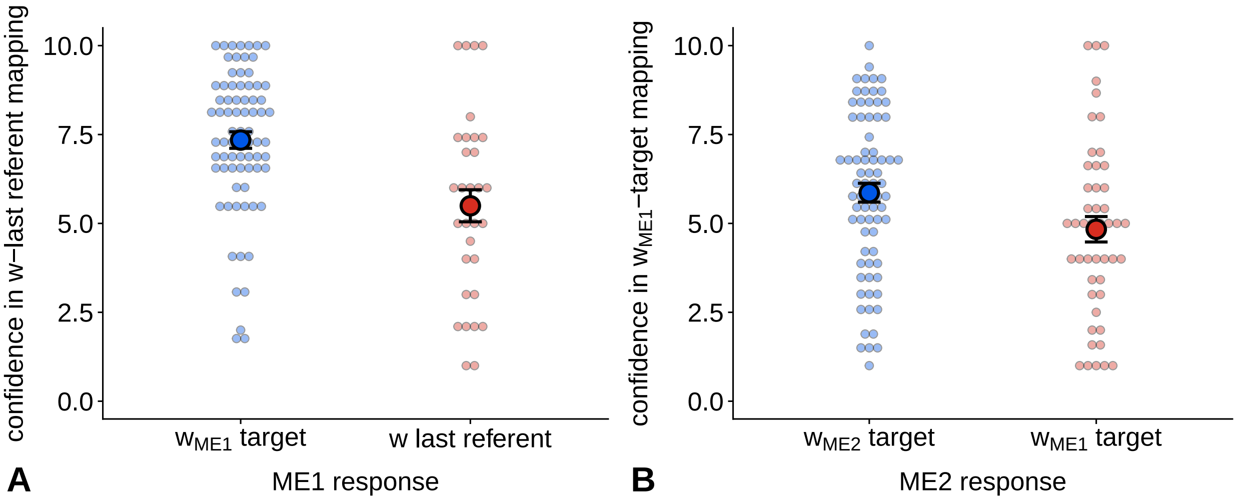

Figure 9. A. Analysis 3 (replication): ME1. Average confidence score in the w-last selected referent mapping as a function of the response in the ME1 trial (w_{ME1} target vs. *w* last referent).**B. Analysis 4 (replication): ME2.** Average confidence score in the w_{ME1}-target mapping as a function of the response in the ME2 trial (w_{ME2} target vs. w_{ME1} target). Dots represent participants' individual response and error bars represent the standard error of the mean.

3.3 Summary of experiment 3

Experiment 3 replicated the findings of experiments 1 and 2. We found consistent evidence of the role of subjective confidence during word learning: First, subjective confidence played a role in the way new information is processed: Holding high confidence in a word-meaning mapping decreased the likelihood that learners would change their mind when observing disconfirming information. Second, subjective confidence in a word-meaning influenced meaning inferences beyond that word: learners were more likely to use a word to make mutual exclusivity inferences if they were confident that they knew the meaning of that word.

4 General Discussion

Across three experiments, we provide evidence that subjective confidence in a word-meaning mapping modulates subsequent information processing in both cross-situational statistical learning and word inference tasks: Learners who were confident they knew the meaning of a word were more likely to persist in their belief than learners who were not, even after observing objective evidence against their belief, and they were also more likely to use that word to learn other words by applying mutual exclusivity.

In theories of word learning, the dominant characterisation has been in terms of accumulation of evidence through learning exposures. In that framework, all information is processed independently of the current state of knowledge of the learner. Our data suggest that this is not what happens: learners can selectively disregard information when they are certain that they already know the meaning of a word. We interpret our findings as showing that subjective confidence influences information selection. However, an alternative explanation is that learning operates through hypothesis sampling (Thaker et al., 2017) which is a more general mechanism that generates hypothesis-testing-like behavior (Stevens et al., 2017; Trueswell et al., 2013): Learners do not maintain all the possible meaning hypotheses, only the ones that are favoured by the data they observed. Concretely, as participants observe learning trials sequentially, the first trials may favour the competitor meaning, while the later trials may favour the correct meaning. If the correct meaning of a word is eliminated during hypothesis sampling during the earliest stages, the early competitor hypothesis may prevail. Such a hypothesis sampling procedure may explain the order effects found here without calling for subjective confidence. This procedure, however, cannot explain the effect of subjective confidence on referent selection while controlling for the objective uncertainty of the trials series, the trial order and participants' past selections. Our data thus strongly favors the interpretation that subjective confidence about a word-meaning shapes both information selection processes and inferential processes during word learning.

In the current cross-situational learning task, where each trial provides new evidence that may help to refine the word-meaning mapping, a bias to persist in one's own belief and to discard new information may seem to be maladaptive. However, in real word learning scenarios, when one's own lexical knowledge is thought to be reliable, it may be rational to ignore conflicting information, presumably outliers or uninformative observations, to increase the robustness of the learning process, especially as the hypothesis space of possible meanings is vast and not explicit (see relatedly Oaksford & Chater, 1994; Qiu, Luu, & Stocker, 2020; Tsetsos et al., 2016). Perhaps most obviously, it is hard to see how inferential processes such as mutual exclusivity would work without assessing the reliability of one's knowledge about word meanings: either one would apply mutual exclusivity immediately, using words for which one does not have much evidence to learn the meaning of other words (which would lead to a cascade of errors), or one would never apply it because it is objectively impossible to eliminate all uncertainty. Subjective confidence appears thus to be a necessary component of the application of mutual exclusivity as it involves using words you think you already know to learn other words, and the current study provides the first direct evidence that subjective confidence distinctively contributes to the application of mutual exclusivity.

It is also important to consider the nature of the word learning process, and the degree to which our findings are particular to how we assessed word learning and probed learners' subjective confidence. In the present study, participants had to make a selection at each trial regardless of their level of confidence in that choice. Learners thus had to explicitly commit to a hypothesis and explicitly introspect about their word knowledge. This was a necessary component of our design as we wanted to track participants hypothesis re-evaluation as a function of their subjective confidence and the objective evidence they were given. This explicit type of learning might overestimate the role of confidence in learning — our task might encourage participants to make use of their confidence ratings, and may simply not reflect how we learn words in the wild or indeed in other cross-situational learning paradigms. Yu et al. (2007), for instance, characterized cross-situational learning as an implicit and automatic process that does not require the learner's awareness. Most tellingly, the authors report that when the experimental task is not presented as a word learning task, and when participants are not required to make a meaning commitment at each trial, participants feel that they did not learn anything but still displayed above-chance performance. Thus, it is possible that our findings may not generalize to the context of implicit word learning, where learners may lack metaknowledge about their knowledge (e.g., Dienes & Berry, 1997). However, the role of subjective confidence may be decisive during intentional reasoning processes that are known to have an important contribution during first language acquisition (Bloom, 2002; M. C. Frank, Goodman, & Tenenbaum, 2009; Tomasello, 2009) or in second language learning settings where lexical knowledge is acquired explicitly.

These considerations relate to an important open question, that our data cannot directly address: whether subjective confidence may influence word learning during children's language acquisition. While adults can readily report on the state of their (linguistic) knowledge, it has been assumed that children's awareness of the reliability of their own linguistic knowledge only emerges at around 4-to-5 years of age, once they have largely acquired their language's structural features, like words and grammar (H. Gleitman & Gleitman, 1979). For instance, children fail to verbally identify novel objects or novel words as ones they don't know until they are about four years old (Marazita & Merriman, 2004; Slocum & Merriman, 2018). However, a recent study provides evidence that even 2-year-old children can evaluate the confidence associated with word recognition (Dautriche, Goupil, Smith, & Rabagliati, 2021), suggesting that even toddlers may be able to use basic forms of metacognition (i.e., "core" metacognition: Goupil & Kouider, 2019; Proust, 2012; Shea et al., 2014), such as the ability to estimate decision confidence, to aid in the process of learning a lexicon. Such confidence estimates could be used by children to optimise how they allocate attention during learning (e.g., ignoring situations in which high-confidence words are used), or to guide interrogative behaviors (e.g., asking clarification when confidence is low), i.e., employing active learning strategies.

Better identifying the role of subjective confidence in word learning has the potential to bridge the gap between formal models of word learning that accumulate evidence through learning instances (Fazly et al., 2010; Xu & Tenenbaum, 2007; Yurovsky & Frank, 2015) and empirical work evidencing active learning behavior in children in the domain of word learning (Bazhydai, Westermann, & Parise, 2020; Hembacher, deMayo, & Frank, 2020; Vaish, Özlem Ece Demir, & Baldwin, 2011; Zettersten & Saffran, 2020). For instance, pre-schoolers preferentially seek information from a social partner when facing ambiguous word-object situations (Bazhydai et al., 2020; Hembacher et al., 2020; Vaish et al., 2011) and actively choose objects whose labels are ambiguous to be given more information about (Zettersten & Saffran, 2020), and adult learners display better word learning accuracy in a cross-situational task when they can choose which word-object associations they want to learn from (Kachergis, Yu, & Shiffrin, 2013). Recent computational models and formal analyses of word learning implement active learning in terms of uncertainty-reduction mechanisms whereby learners attend more to less frequently encountered or more novel object-label associations (Hidaka, Torii, & Kachergis, 2017; Keijser, Gelderloos, & Alishahi, 2019). Yet, our result suggests that learners' evaluation of their (partial) knowledge, rather than the objective state of their partial knowledge, may guide their active learning behavior. This comes with important directions for future research. First, it calls for additional experimental and computational work to disentangle the role of learners' metacognition about their learning from partial knowledge during learning. Second, and critically for the earliest stage of language acquisition, it poses the question of whether learners' *awareness* of their state of knowledge is necessary in guiding selective information processing or whether this could be achieved by core metacognitive processes (Goupil & Kouider, 2019; Proust, 2012; Shea et al., 2014). Finally, an important question concerns the degree to which individual differences in metacognitive abilities could lead to differences in word learning behavior and outcomes. For instance, people with poorer metacognition show less sensitivity to corrective information (Rollwage, Dolan, & Fleming, 2018), suggesting that metacognitive ability might have cascading consequences on the degree to which selective information processing leads to slower vocabulary learning.

To summarize, we showed that subjective confidence in a word-meaning mapping influences word learning. Learners who hold high confidence in a word-meaning mapping were: 1) more likely to ignore conflicting information and 2) more likely to use that word to make mutual exclusivity inferences. We submit that the influence of subjective confidence on word learning should be further explored and incorporated into future accounts of word learning.

References

Baldwin, D. A. (1993). Infants' ability to consult the speaker for clues to word reference. *Journal of child language*, *20*(2), 395–418. doi: 10.1017/S0305000900008345

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Bazhydai, M., Westermann, G., & 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. doi: 10.1111/desc.12938

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual review of psychology*, *64*, 417–444. doi: 10.1146/annurev-psych-113011-143823

Bloom, P. (2002). *How children learn the meanings of words*. MIT press. doi: doi:10.7551/mitpress/3577.001.0001

Carey, S., & Bartlett, E. (1978). Acquiring a Single New Word. *Papers and Reports on Child Language Development*, *15*, 17–29.

Dautriche, I., & Chemla, E. (2014). Cross-situational word learning in the right situations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(3), 892. doi: 10.1037/a0035657

Dautriche, I., Goupil, L., Smith, K., & Rabagliati, H. (2021). Two-year-olds' eye movements reflect confidence in their understanding of words.
doi: 10.31234/osf.io/pd6xh

Desender, K., Boldt, A., & Yeung, N. (2018). Subjective confidence predicts information seeking in decision making. *Psychological science*, *29*(5), 761–778. doi: 10.1177/0956797617744771

Dienes, Z., & Berry, D. (1997). Implicit learning: Below the subjective threshold. *Psychonomic bulletin & review*, *4*(1), 3–23. doi: 10.3758/bf03210769

Diesendruck, G., & Markson, L. (2001). Children's avoidance of lexical overlap: A pragmatic account. *Developmental psychology*, *37*(5), 630. doi: 10.1037/0012-1649.37.5.630

Dobson, A. J., & Barnett, A. G. (2008). *An introduction to generalized linear models*. Chapman and Hall/CRC. doi: 10.1201/9781315182780

Fazly, A., Alishahi, A., & Stevenson, S. (2010). A Probabilistic Computational Model of Cross-

Situational Word Learning. *Cognitive Science*, *34*(6), 1017–1063. doi: 10.1111/j.1551-6709.2010.01104.x

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general bayesian framework for metacognitive computation. *Psychological review*, *124*(1), 91. doi: 10.1037/rev0000045

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*. doi: 10.3389/fnhum.2014.00443

Folke, T., Jacobsen, C., Fleming, S. M., & De Martino, B. (2016). Explicit representation of confidence informs future value-based decisions. *Nature Human Behaviour*, *1*(1), 1–8. doi: 10.1038/s41562-016-0002

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological science*, *20*(5), 578–585. doi: 10.1111/j.1467-9280.2009.02335.x

Frank, S., Goldwater, S., & Keller, F. (2009). Evaluating models of syntactic category acquisition without using a gold standard. In *Proc. 31st Annual Conf. of the Cognitive Science Society* (pp. 2576–2581).

Gleitman, H., & Gleitman, L. (1979). Language use and language judgment. In *Individual differences in language ability and language behavior* (pp. 103–126). Elsevier. doi: 10.1016/c2013-0-10652-6

Gleitman, L. (1990). The Structural Sources of Verb Meanings. *Language Acquisition*, *1*(1), 3–55. doi: 10.1207/s15327817la0101_2

Golinkoff, R. M., Hirsh-Pasek, K., Bailey, L. M., & Wenger, N. R. (1992). Young children and adults use lexical principles to learn new nouns. *Developmental psychology*, *28*(1), 99. doi: 10.1037/0012-1649.28.1.99

Goupil, L., & Kouider, S. (2019). Developing a reflective mind: From core metacognition to explicit self-reflection. *Current Directions in Psychological Science*, *28*(4). doi: 10.1177/0963721419848672

Graham, S. A., Poulin-Dubois, D., & Baker, R. K. (1998). Infants' disambiguation of novel object words. *First Language*, *18*(53), 149–164. doi: 10.1177/014272379801805302

Halberda, J. (2003). The development of a word-learning strategy. *Cognition*, *87*(1), B23–B34. doi: 10.1016/s0010-0277(02)00186-5

Hembacher, E., deMayo, B., & Frank, M. C. (2020). Children's social information seeking is sensitive to referential ambiguity. *Child Development*, *91*(6). doi: 10.1111/cdev.13427

Hidaka, S., Torii, T., & Kachergis, G. (2017). Quantifying the impact of active choice in word learning..

Kachergis, G., Yu, C., & Shiffrin, R. M. (2013). Actively learning object names across ambiguous situations. *Topics in Cognitive Science*, *5*(1), 200–213. doi: 10.1111/tops.12008

Kaplan, J. T., Gimbel, S. I., & Harris, S. (2016). Neural correlates of maintaining one's political beliefs in the face of counterevidence. *Scientific reports*, *6*, 39589. doi: 10.1038/srep39589

Keijser, D., Gelderloos, L., & Alishahi, A. (2019). Curious topics: A curiosity-based model of first language word learning. In *Cogsci* (pp. 1991–1997).

Kiani, R., & Shadlen, M. N. (2009). Representation of confidence associated with a decision by neurons in the parietal cortex. *science*, *324*(5928), 759–764. doi: 10.1126/science.1169405

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, *139*(3), 558. doi: 10.1037/a0019165

Lewis, M., Cristiano, V., Lake, B. M., Kwan, T., & Frank, M. C. (2020). The role of developmental change and linguistic experience in the mutual exclusivity effect. *Cognition*, *198*, 104191. doi: 10.31234/osf.io/wsx3a

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology*, *37*(11), 2098. doi: 10.1037/0022-3514.37.11.2098

Marazita, J. M., & 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. doi: 10.1016/j.jml.2004.06.008

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive psychology*, *20*(2), 121–157. doi: 10.1016/0010-0285(88)90017-5

Medina, T. N., Snedeker, J., Trueswell, J. C., & Gleitman, L. R. (2011). How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, *108*(22), 9014–9019. doi: 10.1093/oso/9780199828098.003.0015

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review*, *15*(1), 174–179. doi: 10.3758/pbr.15.1.174

Meteyard, L., & Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*, 104092. doi: 10.31234/osf.io/h3duq

Meyniel, F., Sigman, M., & Mainen, Z. F. (2015). Confidence as bayesian probability: From neural origins to behavior. *Neuron*, *88*(1), 78–92. doi: 10.1016/j.neuron.2015.09.039

Moss, A. J., Rosenzweig, C., Robinson, J., & Litman, L. (2020). Demographic stability on mechanical turk despite covid-19. *Trends in Cognitive Sciences*. doi: 10.1016/j.tics.2020.05.014

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, *2*(2), 175–220. doi: 10.1037/1089-2680.2.2.175

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608.

O'Hanlon, C. G., & Roberson, D. (2006). Learning in context: Linguistic and attentional constraints on children's color term learning. *Journal of Experimental Child Psychology*, *94*(4), 275–300. doi: 10.1016/j.jecp.2005.11.007

Proust, J. (2012). Metacognition and mindreading: one or two functions? In *Foundations of metacognition* (pp. 234–251). Oxford University Press. doi: 10.1093/acprof:oso/9780199646739.003.0015

Qiu, C., Luu, L., & Stocker, A. A. (2020). Benefits of commitment in hierarchical inference. *Psychological review*, *127*(4), 622. doi: 10.1101/658914

Quine, W. V. (1960). *Word and Object*. Cambridge, MA: MIT. doi: 10.7551/mitpress/9636.001.0001

Rollwage, M., Dolan, R. J., & Fleming, S. M. (2018). Metacognitive failure as a feature of those holding radical beliefs. *Current Biology*, *28*(24), 4014–4021.e8. doi: 10.1016/j.cub.2018.10.053

Rollwage, M., Loosen, A., Hauser, T. U., Moran, R., Dolan, R. J., & Fleming, S. M. (2020). Confidence drives a neural confirmation bias. *Nature communications*, *11*(1), 1–11. doi: 10.32470/ccn.2019.1064-0

Shea, N., Boldt, A., Bang, D., Yeung, N., Heyes, C., & Frith, C. D. (2014). Supra-personal cognitive control and metacognition. *Trends in Cognitive Sciences*, *18*(4), 186–193. doi: 10.1016/j.tics.2014.01.006

Siskind, J. M. (1996). A Computational Study of Cross-Situational Techniques for Learning

Word-to-Meaning Mappings. *Cognition*, *61*(12), 31–91. doi: 10.1016/s0010-0277(96)00728-7

Slocum, J. Y., & Merriman, W. E. (2018, jan). The metacognitive disambiguation effect. *Journal of Cognition and Development*, *19*(1), 87–106. doi: 10.1080/15248372.2017.1415901

Smith, K., Smith, A. D. M., & Blythe, R. A. (2011, April). Cross-Situational Learning: An Experimental Study of Word-Learning Mechanisms. *Cognitive Science*, *35*(3), 480–498. doi: 10.1111/j.1551-6709.2010.01158.x

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568. doi: 10.1016/j.cognition.2007.06.010

Smith, L. B., & Samuelson, L. (2006). An attentional learning account of the shape bias: Reply to Cimpian and Markman (2005) and Booth, Waxman, and Huang (2005). *Developmental Psychology*, *42*(6), 1339–1343. doi: 10.1037/0012-1649.42.6.1339

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(1), 204. doi: 10.1037/0278-7393.26.1.204

Stevens, J. S., Gleitman, L. R., Trueswell, J. C., & Yang, C. (2017). The pursuit of word meanings. *Cognitive science*, *41*, 638–676. doi: 10.1111/cogs.12416

Talluri, B. C., Urai, A. E., Tsetsos, K., Usher, M., & Donner, T. H. (2018). Confirmation bias through selective overweighting of choice-consistent evidence. *Current Biology*, *28*(19), 3128–3135. doi: 10.1016/j.cub.2018.07.052

Thaker, P., Tenenbaum, J. B., & Gershman, S. J. (2017, April). Online learning of symbolic concepts. *Journal of Mathematical Psychology*, *77*, 10–20. doi: 10.1016/j.jmp.2017.01.002

Tillman, K. A., & Barner, D. (2015). Learning the language of time: Children's acquisition of duration words. *Cognitive psychology*, *78*, 57–77. doi: 10.31234/osf.io/5afc3

Tomasello, M. (2009). *Constructing a language: A usage-based theory of language acquisition*. Harvard university press.

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013, February). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*(1), 126–156. doi: 10.1016/j.cogpsych.2012.10.001

Tsetsos, K., Moran, R., Moreland, J., Chater, N., Usher, M., & Summerfield, C. (2016). Economic irrationality is optimal during noisy decision making. *Proceedings of the National Academy of*

Sciences, *113*(11), 3102–3107. doi: 10.1073/pnas.1519157113

Vaish, A., Özlem Ece Demir, & Baldwin, D. (2011). Thirteen- and 18-month-old infants recognize when they need referential information. *Social Development*, *20*(3), 431–449. doi: 10.1111/j.1467-9507.2010.00601.x

Waxman, S. R., & Klibanoff, R. S. (2000). The role of comparison in the extension of novel adjectives. *Developmental psychology*, *36*(5), 571. doi: 10.1037/0012-1649.36.5.571

Xu, F., Cote, M., & Baker, A. (2005). Labeling guides object individuation in 12-month-old infants. *Psychological Science*, *16*(5), 372–377.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272. doi: 10.1037/0033-295X.114.2.245

Yang, C. (2020). How to make the most out of very little. *Topics in Cognitive Science*, *12*(1), 136–152. doi: 10.1111/tops.12415

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1310–1321. doi: 10.1098/rstb.2011.0416

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, *18*(5), 414. doi: 10.1111/j.1467-9280.2007.01915.x

Yu, C., Smith, L. B., Klein, K., & Shiffrin, R. M. (2007). Hypothesis testing and associative learning in cross-situational word learning: Are they one and the same. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 737–742).

Yurovsky, D., & Frank, M. C. (2015). An integrative account of constraints on cross-situational learning. *Cognition*, *145*, 53–62. doi: 10.1016/j.cognition.2015.07.013

Yurovsky, D., & Yu, C. (2008). Mutual exclusivity in cross-situational statistical learning. In *Proceedings of the 30th annual conference of the Cognitive Science Society* (pp. 715–720).

Yurovsky, D., Yu, C., & Smith, L. B. (2013). Competitive processes in cross-situational word learning. *Cognitive Science*, *37*(5), 891–921. doi: 10.1111/cogs.12035

Zettersten, M., & Saffran, J. R. (2020, dec). Sampling to learn words: Adults and children sample words that reduce referential ambiguity. *Developmental Science*, *24*(3). Retrieved from https://doi.org/10.1111%2Fdesc.13064 doi: 10.1111/desc.13064