# Expectation Violation Enhances the Development of New Abstract Syntactic Representations: Evidence from an Artificial Language Learning Study

Giulia Bovolenta
Emma Marsden
University of York, UK

**Abstract:** Prediction error is known to enhance priming effects for familiar syntactic structures; it also strengthens the formation of new declarative memories. Here, we investigate whether violating expectations may aid the acquisition of new abstract syntactic structures, too, by enhancing memory for individual instances which can then form the basis for abstraction. In a cross-situational artificial language learning paradigm, participants were exposed to novel syntactic structures in ways that either violated their expectations (Surprisal group) or that conformed to them (Control group). First, we established a potential expectation to hear feedback that simply repeated the same structure as that just experienced. We then manipulated feedback so that the Surprisal group unexpectedly heard passive structures in feedback following active sentences, while the Control group only heard passive structures following passive sentences. Delayed post-tests examined participants' structural knowledge both by means of structure test trials (focusing on the active / passive distinction, with both familiar and novel verbs), and by a grammaticality judgment task. The Surprisal group was significantly more accurate than the Control group on the structure test trials with novel verbs and on the grammaticality judgment task, suggesting participants had developed stronger abstract structural knowledge and were better at generalising it to novel instances. Tentative evidence suggested the Surprisal group was not significantly more likely to become aware of the functional distinction between the two structures.

**Keywords:** Second language acquisition; prediction error; cross-situational learning; artificial language learning; structural generalisation.

**Corresponding author(s):** Giulia Bovolenta, Department of Education, University of York, Heslington, York, YO10 5DD, UK. Email: giulia.bovolenta@york.ac.uk.

**ORCID ID(s):** Giulia Bovolenta: https://orcid.org/0000-0003-4139-6446;
Emma Marsden: https://orcid.org/0000-0003-4086-5765.

**Citation:** Bovolenta, G., & Marsden, E. (2021). Expectation Violation Enhances the Development of New Abstract Syntactic Representations: Evidence from an Artificial Language Learning Study. *Language Development Research*, DOI: 10.34842/c7t4-pz50

## Introduction

Is it possible to 'surprise' a learner into acquiring a new structure in a foreign language? A growing body of literature suggests that unpredictable input favours language learning. On one hand, structural adaptation – an increased likelihood to use or expect the syntactic structures we are exposed to, persisting in the long term – is likely one of the mechanisms by which we tune into the patterns of our language (Peter & Rowland, 2019). There is evidence that prediction error drives adaptation to syntactic structure, both from computational modelling (Chang, Dell, & Bock, 2006) and empirical studies with both first language (L1) and second language (L2) speakers (Fazekas, Jessop, Pine, & Rowland, 2020; Montero-Melis & Jaeger, 2020). At the same time, evidence shows that violating expectations facilitates the formation of new individual declarative memories, too, including vocabulary learning (Greve, Cooper, Kaula, Anderson, & Henson, 2017; Stahl & Feigenson, 2017). We are now beginning to form a picture of the ways in which surprisal can aid learning with regards to different aspects of language. If a learner already has the relevant abstract syntactic representation, encountering the structure in a surprising context appears to strengthen that representation. Surprisal can also facilitate the acquisition of new declarative memories for lexical items, such as nouns or verbs, leading to stronger memory formation than non-surprising contexts. But what about the acquisition of *new, syntactic* representations among adult learners who have already established their L1 system? In this study, we address an unexplored gap in the literature, asking whether surprisal could also aid the development of new abstract structural representations, including acquisition of their specific form-meaning mappings, rather than just strengthening existing ones. Following a usage-based approach to language acquisition, we assume that structural knowledge emerges through abstraction from individual learned exemplars (N. C. Ellis, Römer, & O'Donnell, 2016). If expectation violation can aid memory for individual instances, then we hypothesise that it may also aid the acquisition of structural knowledge through abstraction from these individual instances.

We investigated this question in a controlled learning experiment using an artificial language (Yorwegian). Learners were first introduced to a default syntactic structure, the active construction, which they learned while they were also learning the vocabulary of the language. Then, once this structure had been learned and consolidated, participants were exposed on the second day to a (potentially) more complex alternative, the passive construction. This ordering (active then passive) and bias in the input (more active than passive) simulates, to some extent, the likely real-life learning experience of many learners, who would tend to encounter the passive construction less often in their learning due to its lower frequency, relative to the active construction. In this context, we manipulated the utterance containing the passive construction (in what we called a 'feedback' turn), so as to make it either unexpected (Surprisal group) or expected (Control group) relative to the pattern that had been established during training. Participants responded to sentences they heard (by selecting the matching picture) and received feedback on their responses, which consisted of a replay of (the meaning of) the initial sentence they had given their response to. In the first blocks, both groups received feedback using the structure that was always congruent with the structure in the initial sentence, i.e., participants heard the exact same sentence. However, in later trials, the Surprisal group

occasionally experienced feedback containing a passive structure immediately following an active structure (though still describing the same picture and with the same meaning in terms of agents and patients), while the Control group always experienced feedback containing the structure that matched the one used in the preceding sentence. We hypothesised that participants in the Surprisal group would develop stronger representations for the passive sentences encountered in feedback, leading to improved learning of the passive syntactic structure itself[1]. In a secondary question, we also hypothesised that surprisal may aid the development of explicit knowledge, either by increasing attention and cognitive effort (Leow, 2015) or by generating stronger representations that would be more likely to emerge in conscious awareness (Cleeremans, 2011).

## Background Literature

### Structural Priming as a Learning Mechanism

When language users encounter a particular syntactic construction, they are often more likely to expect it again, or to use it in production, than they were before encountering it, a phenomenon known as *structural priming* (Arai, van Gompel, & Scheepers, 2007; Bock, 1986; Ferreira & Bock, 2006; Ledoux, Traxler, & Swaab, 2007). When the priming effect persists over time, it is known as *adaptation* (Kaan & Chun, 2018b). Adaptation to syntactic structure alternations (such as that between prepositional object and double object dative constructions in English) has been observed in L1 production (Jaeger & Snider, 2013; Kaschak, 2007; Kaschak & Borreggine, 2008; Kaschak, Kutta, & Jones, 2011; Kaschak, Loney, & Borreggine, 2006), and in L1 comprehension (Farmer, Fine, Yan, Cheimariou, & Jaeger, 2014; Fine & Jaeger, 2016; Fine, Jaeger, Farmer, & Qian, 2013; Kaan & Chun, 2018a). Adaptation effects have also frequently been observed in L2 speakers (Jackson & Ruf, 2017; Kaan & Chun, 2018a; McDonough & Trofimovich, 2015; Montero-Melis & Jaeger, 2020; Shin & Christianson, 2012; see Jackson, 2018 for a review). The magnitude of these effects tends to be greater for less frequent structures (known as *inverse probability effects*). This has been observed empirically in both the L1 and L2: Structures that have lower frequency in the input elicit greater priming effects (Hartsuiker, Kolk, & Huiskamp,

[1] In this sense, our manipulation is quite different from previous research on surprisal in language processing, as it manipulates expectations about the context in which the 'surprising' language was experienced, rather than the input per se. A reviewer pointed out that another potential way of framing our manipulation could perhaps be as a type of 'recast', which is an interactional and/or feedback (error correction) phenomenon, both in natural discourse (e.g., as a confirmatory turn or as a clarification/comprehension checking mechanism) and in language instruction (e.g., confirmatory to promote continued communication, or corrective to provide feedback on errors) (Goo & Mackey, 2013; Lyster & Saito, 2010; see, however, Foster (Foster, 1998; Foster & Ohta, 2005), who downplays the frequency of recasting in instructional situations). In our study, an incongruent (potentially conceptualised as 'corrective' or 'comprehension checking') recast could be more salient and/or lead to greater awareness relative to a congruent (potentially conceptualised as 'confirmatory' or 'interaction promoting') one, a possibility we raise in the discussion. However, one caveat to keep in mind is that recasts in L2 acquisition studies are normally in response to an utterance produced by the learner, whereas in our study the initial statement is *heard* by the learner, rather than produced. Therefore, if we think of our study in terms of recast, our design could perhaps simulate the cases where a learner hears the interaction and is working out the meaning, rather than actively participate in the interaction.

1999; Hartsuiker & Westenberg, 2000; Jaeger & Snider, 2013; Kaan & Chun, 2018a; Kaschak, 2007; Kaschak et al., 2006; Montero-Melis & Jaeger, 2020; Weber, Christiansen, Indefrey, & Hagoort, 2019).

In L2 acquisition, there is evidence of syntactic priming mechanisms operating from the earliest stages of learning. In Weber et al. (2019), participants were exposed to a novel artificial language in four sessions over the course of nine days. The language consisted of a lexis containing novel nouns and verbs arranged in four possible word orders: three transitive (VOS, OSV, SVO) and one intransitive (SV). In the first session, participants were pre-trained on the nouns. From the second session, participants read novel sentences aloud, which were accompanied by pictures depicting their meaning. Repetition of word order and verb was manipulated from one trial to the next to study priming effects, which were measured by read-aloud times. In the third and fourth session, priming was also assessed through a picture matching task after target trials, where participants had to pick the picture matching the sentence they had just read, out of two possible alternatives (the incorrect picture depicted the same event, but with Agent and Patient roles reversed). Priming effects in read-aloud times were observed from the earliest stages; however, there was no difference in magnitude of priming for infrequent vs. frequent structures (structure frequency was manipulated in the second session, where the frequent word order was twice as likely to occur as the other three). Priming effects were also observed in comprehension, with higher accuracy for repeated structures, but only if they were the frequent word order.

It has been suggested that structural priming is a case of implicit error-based learning (Bock, Dell, Chang, & Onishi, 2007; Bock & Griffin, 2000; Chang, Janciauskas, & Fitz, 2012). Computational modelling of priming data shows that this can be reproduced by a Recurrent Neural Network (RNN) model trained on next-word prediction. As the model encounters more sentences, it gradually improves by adjusting its predictions based on the discrepancy between predicted and actual input, or *prediction error* (Chang et al., 2006). This account is compatible with observed properties of priming and adaptation, such as inverse frequency effects: Low-frequency words would generate greater prediction error, causing a larger adjustment in the weights and therefore a larger learning effect (Chang et al., 2006). More recently, Fazekas et al. (2020) tested this account with an empirical study with both adults and children, and found that exposing participants to surprising dative sentences (using verbs rarely associated with the dative structure) made participants more likely to use the dative structure in a post-test.

**Effect of Expectation Violation on New Memory Formation**

Structural priming and adaptation phenomena affect representations that have *already* been acquired; what changes as a consequence of exposure, and is further increased by prediction error, is the strength of existing structural representations. However, evidence from a different strand of research, originating mainly in cognitive psychology, shows that prediction error can also enhance the formation of *new* individual memories; events or associations which violate our expectations are remembered better than those that conform with them (*one-shot declarative learning*). Novel associations are better remembered if they violate an established pattern

(Brod, Hasselhorn, & Bunge, 2018; Greve et al., 2017; Greve, Cooper, Tibon, & Henson, 2019), including translation word pairs (De Loof et al., 2018). Surprising feedback, too, is better remembered. Fazio & Marsh's (2009) participants answered general knowledge questions (rating their confidence in their answers) and then were shown the correct answer, which was displayed in either red or green letters. When feedback was unexpected (either following a high-confidence incorrect answer, or a low-confidence correct one) memory for the font colour in which it was displayed was better than for expected feedback. This suggests that surprising feedback can lead to a greater effort to encode it (known as the *surprise hypothesis),* resulting in better 'source memory' (defined as memory for the conditions in which the feedback is encoded, including everything that gets encoded besides the content of the feedback itself).

There is also direct evidence that the effect of violation expectation on novel memory formation can aid language acquisition: Stahl & Feigenson (2017) showed that violation of expectations promotes vocabulary learning in young children. In the study, 3- to 6-year-old children were exposed to novel events which were either entirely possible or which violated core properties of the objects involved (e.g., a cup vanishing and reappearing in a different location). They were then taught the verb corresponding to the action (Experiment 1) or the noun denoting one of the objects (Experiment 2), and were tested immediately on its meaning. Children were significantly more accurate in their responses for verbs and nouns that they had learned in surprising events than for those they had learned in expected events (on which they performed at chance level). The effect was limited to nouns and actions involved in the surprising event: If children were taught the name for an object that was present during the event but did not participate in it, there was no learning effect (Experiment 4). This suggests that violated expectation did not aid learning simply by increasing attention or arousal, but that it led children to revise their predictions about specific objects and events (Stahl & Feigenson, 2017).

## The Present Study

From the literature surveyed, it is clear that unexpected input can lead to a strengthening of existing abstract structural representations, in the form of increased priming and adaptation. We also know that violated expectation enhance the formation of new declarative memories, including learning novel vocabulary items. What we do not know, and what is the of focus of this study, is whether surprisal may also favour the acquisition of new *abstract structural* representations. In usage-based accounts of language acquisition, the development of abstract, structural knowledge is assumed to proceed from learned exemplars in the first place (Bybee & Hopper, 2001; N. C. Ellis, 2002; N. C. Ellis et al., 2016). If expectation violation can aid memory for individual instances, then we hypothesise that it may also aid the acquisition of structural knowledge through abstraction from these individual instances.

For our study, we adapted a cross-situational learning paradigm (Smith & Yu, 2008; Yu & Smith, 2007) which has been successfully used in previous studies to investigate the acquisition of syntax in naturalistic settings (Monaghan, Ruiz, & Rebuschat, 2020; Rebuschat, Monaghan, & Schoetensack, 2021; Walker, Monaghan, Schoetensack, &

Rebuschat, 2020). In a cross-situational learning paradigm, participants are exposed to a novel language without any explicit instruction, but instead derive the meaning of novel words by attempting to interpret them across multiple situations. Participants are exposed to novel words or sentences and are required to select the correct interpretation from a range of options. While they are initially at chance in their answers, participants eventually converge on the correct meaning by keeping track of possible interpretations across different trials. Walker et al. (2020) used an artificial language composed of 16 novel words (8 nouns, 4 verbs, 2 adjectives) which could be arranged in either a subject-object-verb (SOV) or object-subject-verb (OSV) word order. Participants were trained and tested on the language over the course of two days, without any explicit instruction. In each learning trial, they heard a sentence in the novel language while two animations appeared on screen; their task was to select the one matching the sentence.

Accuracy in learning trials was above chance from the second block, and results from intermitting test blocks showed that participants succeeded in acquiring both the grammar and vocabulary. This makes it a highly suitable paradigm to investigate the acquisition of syntactic structure in a naturalistic way. To establish expectation and then induce surprisal, we added feedback to critical trials. This feedback always contained a passive structure, which was, at first, always consistent with the trial just heard. We then manipulated the feedback between groups to be either consistent or inconsistent with expectations that participants had established during their first blocks of feedback trials. That is, we assumed that participants would expect feedback turns to replay the sentence in the exact form they had just heard. To generate this expectation, we ensured that feedback was initially congruent for both groups, and only at a later stage did we introduce, for the Surprisal group only, incongruent trials: active sentences that were followed by a passive form, whilst the *same* picture was displayed as during the active sentence. Given that surprising feedback is thought to be better encoded, including its visual features (Fazio & Marsh, 2009), we expected the passive sentences in surprising feedback trials to lead to better learning, not only of the picture itself, but of the specific sentence – picture pairing, too, relative to the learning in the group that experienced the expected feedback trials.

It is also possible that surprising feedback may promote the development of explicit knowledge of the passive structure. While findings like those of Stahl & Feigenson (2017) suggest that the effect of expectation violation on learning is not driven simply by a *general* raising of attention, it seems likely that surprisal has an effect on attention, albeit only to the relevant features (see for instance Greve et al. (2017) on possible mechanisms underlying one-shot declarative learning). In the context of associative learning, it has been suggested that surprisal may increase the salience of a stimulus, which in turn drives learning (Cintrón-Valentín & Ellis, 2016; N. C. Ellis, 2016, 2017). Increased attention may also lead to greater awareness, that is, explicit knowledge of the form-meaning connections being learned. On the one hand, this may happen directly as a consequence of deeper engagement with the stimuli; for example, in L2 research, greater cognitive effort has been reported to correlate positively with the emergence of rule awareness (Cerezo, Caras, & Leow, 2016; Leow, 2015). On the other hand, surprisal may also have a more indirect effect on the emergence of explicit knowledge. According to the *radical plasticity* thesis

(Cleeremans, 2008, 2011), there is a continuum between implicit and explicit knowledge. On initial exposure, implicit knowledge develops, characterised by weak and low-quality representations in memory. As the quality and strength of representation increase with repeated exposure, the knowledge becomes increasingly available to consciousness, that is, becomes explicit. Therefore, if surprisal leads to stronger representations in memory, we may also expect it to lead to more explicit knowledge. More specifically, in our case, stronger representations of individual passive sentences may lead to greater awareness of the form-meaning connections involved (though we do not aim to tease apart these two accounts [greater cognitive effort versus radical plasticity] of how this may happen).

**Research Questions and Predictions**

Our primary research question (RQ1) was whether being exposed to surprising items in the passive would lead to overall better knowledge of the passive structure. This was assessed by performance accuracy on picture-matching comprehension tests, with both trained and novel lexicon (to assess generalisation to new instances), and a grammaticality judgment task. If expectation violation can aid structural learning, we would expect the Surprisal group (SG) to show better knowledge of the passive structure than the Control group (CG).

Specifically, with regards to comprehension (in picture-matching comprehension tests), we predicted the Surprisal group would perform better than the Control group in structure comprehension test blocks which were placed both at the end of Day 2, after the surprisal manipulation was introduced, and on Day 3. Day 3 included structure test using both previously trained and novel verbs; we expected the Surprisal group to perform better than Control on both tests. Additionally, we introduced individual comprehension test trials on Day 2 immediately after surprising items, to test for any immediate effects of surprisal on structure comprehension. If surprisal led to increased priming effects, too, we would expect the Surprisal group to perform better than the Control group in structure comprehension immediately after surprising passive items. In all comprehension tests (blocks and individual trials) our prediction of an advantage for the Surprisal group concerned the passive structure only, given that this was the structure affected by the surprisal manipulation. We did not expect to observe any effects on the active structure. In the Grammaticality Judgment Task, too, we expected the Surprisal group to perform better than the Control group in their ability to correctly discriminate between grammatical and ungrammatical passive sentences. We did not expect to see any significant differences between groups in their ability to discriminate between grammatical and ungrammatical sentences in the active form. Our secondary research question (RQ2) concerned the possible effects of surprisal on the development of explicit knowledge. Explicit knowledge of the novel structures was assessed by retrospective verbal report, with a debriefing questionnaire administered at the end of study. If expectation violation can promote the development of explicit knowledge, we would expect the SG to show higher rates of awareness than the CG.

This was the first study of a planned project involving data collection from different populations, both online and in the laboratory. Therefore, we also collected a set of

cognitive measures (procedural learning abilities and verbal declarative memory) which mediated performance in a previous cross-situational learning study (Walker et al., 2020) in order to control for potential effects of individual differences. However, we did not have any specific predictions regarding possible interactions between the individual differences we measured and the surprisal manipulation.

## Methods

### Participants

To our knowledge, there are no previous studies that attempted to investigate the effect of expectation violation on structural learning. This means that we had no single point of reference we could use to estimate the potential size of the effect we were interested in, in order to determine a suitable sample size. Therefore, we based our group size calculations on a set of previous studies each investigating one aspect of our manipulations. Walker et al. (2020), from whom we adapted the cross-situational learning paradigm, tested two groups of 32 subjects, which was estimated by the authors to give .99 power for the simultaneous acquisition of two syntactic structures, based on the effect size from their previous study using the same paradigm. Greve et al. (2017), who investigated the effect of prediction error when learning for novel picture-word associations, used a range of group sizes from 20 to 36 subjects, the latter of which was calculated to have .75 power for one-shot declarative learning.

The effect we were interested in was the interaction between these two aspects, namely the effect of surprisal *on* the acquisition of syntactic structure. However, we had no means of estimating the effect size of a potential interaction. Therefore, we designed the study to have at least enough power to detect the two effects separately, on the assumption that this would be a necessary (although not necessarily sufficient) condition to detect the interaction, if one existed. Based on these considerations, we estimated that a group size of at least 35 participants would be the minimum sample size that we should use in the study.

76 native speakers of English (59 females, $M_{AGE}$ = 31, $SD$ = 7.62) were recruited via the online research platform Prolific (https://www.prolific.co/) and completed the study over the course of three consecutive days, receiving compensation of £12. The study was given ethics approval by the Education Ethics Committee at the University of York. Participants all reported living in the United Kingdom at the time of taking part in the study. Only one participant reported knowledge of any Scandinavian language (Norwegian) (upon which our artificial language was based); this was at the beginner level and they stated that they had never received formal instruction in the language. Participants were randomly assigned to either the Surprisal (n = 39) or Control (n = 37) group on the first day of the study. The slight numerical imbalance between groups is a consequence of attrition (i.e., participants were evenly assigned to the two conditions on Day 1, but not all completed all three days).

## Materials

All stimuli and experimental scripts can be downloaded from the OSF repository for this study (https://doi.org/10.17605/OSF.IO/NKSU8) and from the IRIS database (https://www.iris-database.org/). Participants were trained in an artificial language called Yorwegian, consisting of four nouns (*glim, blom, prag, meeb* – man, woman, boy, girl), eight verbs (*flug-, loom-, gram-, pod-, zal-, shen-, norg-, klig-* – call, chase, greet, interview, pay, photograph, scare, and threaten), one determiner (*lu* - the) and one preposition (*ka* - by). Some of the lexical items used were adapted from Wonnacott, Newport, & Tanenhaus (2008). The specific word-meaning pairs within the noun and verb categories were randomly assigned for every participant. All sentences were SVO, but there were two possible syntactic structures, differentiated by verbal inflection and use of the preposition *ka*. These were the Active structure (e.g. *Lu meeb flug**at** lu prag,* 'The girl calls the boy') and the Passive (e.g. *Lu prag flug**es** **ka** lu meeb,* 'The boy is called by the girl'). This type of passive construction is naturally found in Scandinavian languages. It was chosen so as to have a way of forming passive structures that would not be entirely familiar to L1 English speakers (as there is no equivalent of the BE auxiliary in Yorwegian), while still being ecologically valid.

We used a set of 208 black and white photographs depicting transitive actions, which we adapted from materials created by Segaert and colleagues (Menenti, Gierhan, Segaert, & Hagoort, 2011; Segaert, Menenti, Weber, Petersson, & Hagoort, 2012). The main set of training and testing pictures used on all three days (192 images) depicted the eight verbs: *call, chase, greet, interview, pay, photograph, scare,* and *threaten*. There were four characters which could fill the roles of Agent and Patient: *man, woman, girl* and *boy*. All possible combinations of different characters were included for each training verb, which yielded 12 possible Agent-Patient combinations (the Agent and Patient were always played by different characters). In the training set, the 12 Agent-Patient combinations were repeated for each of the eight verbs, yielding a total number of 96 possible scenes. Each scene was enacted twice, each with different actors, giving a total of 192 unique training pictures. Each picture could appear with one of two possible syntactic structures (Active and Passive constructions), for a total of 384 unique picture-sentence combinations.

The first 96 training pictures (Actor set 1) were used for training blocks on Day 1 and then again on Day 2. On Day 1, all training pictures appeared in the Active construction; on Day 2, half were presented in the Active, the other half in the Passive construction (which pictures appeared in each structure was counterbalanced across participants). Pictures from the second half (Actor set 2) were used for testing blocks distributed across the three days: vocabulary testing blocks on Day 1, Day 2, and Day 3, as well as structure test blocks on Day 2 and Day 3. No unique picture-sentence combination was presented more than once over the course of the experiment. An additional 'generalisation set' was also used (16 images). The pictures in this set depicted four additional transitive verbs (*dress, hug, pull,* and *push*) and were used in a generalisation structure test block on Day 3, to test participants' ability to process the syntactic structures they had been previously exposed to when used with novel verbs. This set used a reduced number of Agent-Patient combinations (four in total: *man-woman, woman-man, boy-girl, girl-boy*).

## Procedure

Participants took part in the study online over the course of three consecutive days ( Figure 1). The average total duration of the study was ~75 min, with each of the three sessions taking approximately 25 min. On each day, participants had to complete the session between 10am and 6pm. Subjects were randomly assigned to one of two groups, Surprisal or Control. On Day 1, the two groups followed the exact same protocol. On Day 2, all participants followed the same procedure in blocks 1-4. In blocks 2 and 3, feedback was introduced and was the same for both groups. In blocks 5-8, we introduced the between-group surprisal manipulation (described in the next section, on 'Learning trials with feedback'). On Day 3, both groups again followed the same protocol throughout. Participants performed the main task (the cross-situational learning paradigm) over the course of three days. On Day 3, this was followed by a grammaticality judgment task, a serial reaction task, the LLAMA B3 test, and a debriefing questionnaire. All tasks were created using JavaScript library PsychoJS, based on PsychoPy (Peirce et al., 2019), with the exception of the LLAMA B3 test, which was built in jsPsych (De Leeuw, 2015). All experimental scripts were hosted and run online through platform Pavlovia (https://pavlovia.org/). Surveys at the end of the experiment were administered using Qualtrics (www.qualtrics.com).

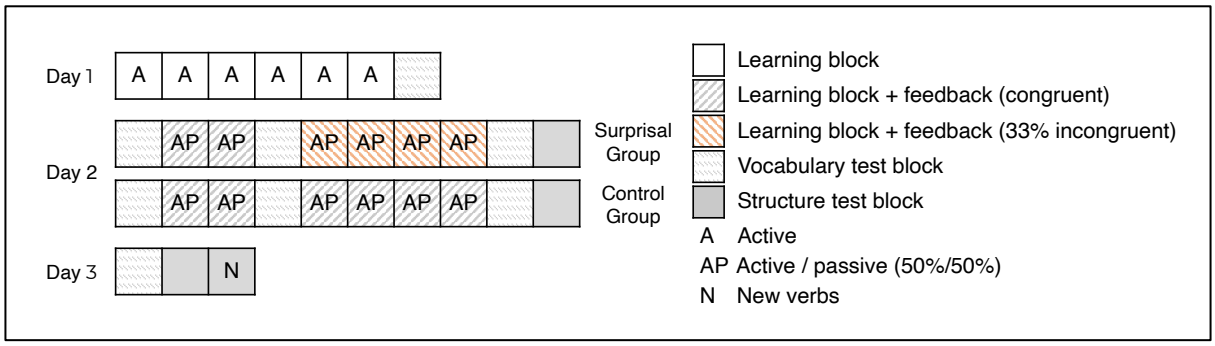


**Figure 1.** ***Summary of cross-learning task schedule***

### *Cross-situational Learning Task*

Participants received no explicit instruction on either the grammar rules or vocabulary of Yorwegian. They were taught using an adapted version of the cross-situational task used by Walker et al. (2020), which was also used for testing. Participants heard individual sentences in Yorwegian, while two pictures (a target picture and a distractor picture) appeared on screen side by side. Their task was to select the picture that corresponded to the sentence they just heard (the target) by pressing the left or right arrow on their keyboard. There were four different types of trials: normal learning trials, vocabulary test trials, structure test trials, and learning trials with feedback (which included the critical between-group manipulation). In normal learning and testing trials, participants received no feedback on their answers.

*Normal learning trials.* Distractor Agent, Patient, and verb were picked by the experimental software at random, with the only constraint being that the distractor

verb could not be the same as the target verb (to avoid the possibility of participants seeing two pictures depicting the same scene, only enacted by different actors).

*Learning trials with feedback.* On Day 2, all learning blocks (Blocks 2-3 and 5-8), contained a proportion of learning trials with feedback. 12 out of 16 learning trials in each of these blocks were followed by feedback on the answer just given: after making their choice (in a learning trial), participants were shown the correct picture which they should have picked, regardless of whether they had picked it or not (in a feedback screen). They saw the correct picture displayed on its own, in the centre of the screen, and they also heard the sentence which they had responded to once again. More precisely, they heard a sentence with the same Agent, Patient and verb as the one they had responded to, but, depending on the block and group, the syntactic structure used to describe the scene could either be the same (congruent feedback) or different (incongruent). In Blocks 2 – 3, all feedback was congruent and evenly spread across structures: both groups received feedback on 6 passive and 6 active learning trials per block, and the sentence they heard during feedback matched the one they had responded to, in both content (meaning) and structure. This was done to ensure that both groups would develop an expectation for feedback to replay sentences using the same structure.

**Table 1.** ***Types of trial included in critical learning blocks (Blocks 5 - 8). Differences between groups are highlighted in bold.***

| Group | Feedback | Main structure (heard first) | Feedback structure | Type | n |
|---|---|---|---|---|---|
| Control | No | **Active** | - | No feedback | 4 |
| Control | Yes | Active | Active | Congruent | 4 |
| Control | Yes | Passive | Passive | Congruent | 8 |
| Surprisal | No | **Passive** | - | No feedback | 4 |
| Surprisal | Yes | Active | Active | Congruent | 4 |
| Surprisal | Yes | Passive | Passive | Congruent | 4 |
| Surprisal | Yes | **Active** | **Passive** | **Incongruent** | 4 |

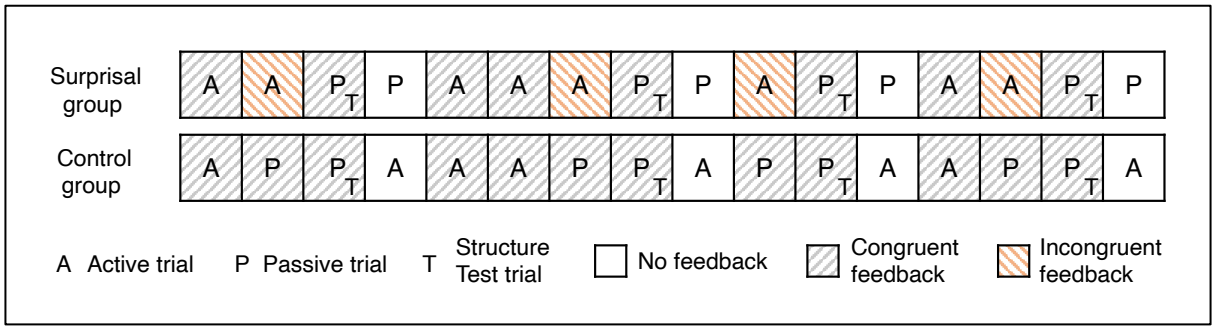


**Figure 2. Example of a critical learning block (Blocks 5-8)**

In Blocks 5 – 8, we introduced the between-group 'surprisal' manipulation. Feedback was still given on 12 out of 16 trials, and both groups still received congruent feedback

on 8 of these 12 trials (4 active and 4 passive). The remaining 4 learning trials with feedback were manipulated so that the feedback they were followed by was congruent for the Control group, but incongruent for the Surprisal group (Table 1 & Figure 2). The only difference between congruent and incongruent feedback was the structure used: in both cases, the correct picture was shown, and the sentence which was heard had the same meaning as that heard during training. However, in incongruent trials the sentence was recast in the opposite syntactic structure (which was the 'incongruent' aspect), while in congruent feedback the exact same sentence was re-played, both with regards to meaning and syntactic structure used. In the Control group, these 4 critical trials required participants to respond to a passive sentence, while in the Surprisal group participants would respond to an active one. This was done to ensure that the feedback itself – the sentence learners were exposed after giving their answer, as they saw the correct picture again – would be in the passive for both groups. This manipulation meant that 8 of 12 trials with feedback used an active structure for the Control group, while for the Surprisal only 4 out of 12 were in the active form. To compensate for this imbalance and ensure the same amount of exposure to the structures in both groups, the remaining 4 trials in each block (which did not have feedback) were manipulated to be passive for the Surprisal group, and active for the Control group (Figure 2). Over the course of the whole experiment, participants saw 16 critical learning trials with feedback (with incongruent feedback for the Surprisal group, but congruent for Control), four in each of Blocks 5 to 8. Each of these critical trials was followed by a structure test trial, which is described below.

*Structure test trials.* All parameters in the pictures were kept constant apart from the Agent and Patient roles, which were reversed from target to distractor pictures (e.g., if the target picture was *The girl interviews the man*, the distractor would be *The man interviews the girl*). Distractor pictures were always picked randomly from either Actor set 1 or 2, regardless of which Actor set the target picture was drawn from (this was done to increase engagement and avoid creating a sense that there was any difference between blocks, which would have been the case if individual blocks only ever showed pictures from one particular set). The following parameters were always randomly chosen: the position of target and distractor picture on screen (left / right), and the position of Agent and Patient characters inside the pictures (left / right). Structure test trials were included in structure test blocks and also immediately following critical feedback trials.

*Noun test trials.* All parameters in the pictures were kept constant apart from the Patient noun (e.g., if the target picture was *The girl interviews the man*, the distractor could be *The girl interviews the boy* or *The girl interviews the woman*). Noun test trials were included in vocabulary test blocks only.

*Verb test trials.* All parameters in the pictures were kept constant apart from the verb. Verb test trials were included in vocabulary test blocks only.

### *Auditory Grammaticality Judgment Task*

Following the cross-situational learning task on Day 3, participants did an auditory grammaticality judgment task (a widely used technique – see Plonsky, Marsden,

Crowther, Gass, & Spinner (2020) with novel Yorwegian sentences. They were instructed to listen to each sentence and indicate whether it was a correct sentence in the language they had been learning. After each sentence was played, the words CORRECT and INCORRECT appeared side by side on screen, and participants had to press either the left or right arrow on their keyboard to give a response. Responses were untimed and ample time was given to respond; the next sentence was shown only after participants gave a response. They heard a total of 32 sentences, 16 grammatical and 16 ungrammatical. Half of the ungrammatical sentences contained the active verbal inflection followed by the agent marker, while the other had a passive verb, but no agent marker (see Table 2 for example stimuli).

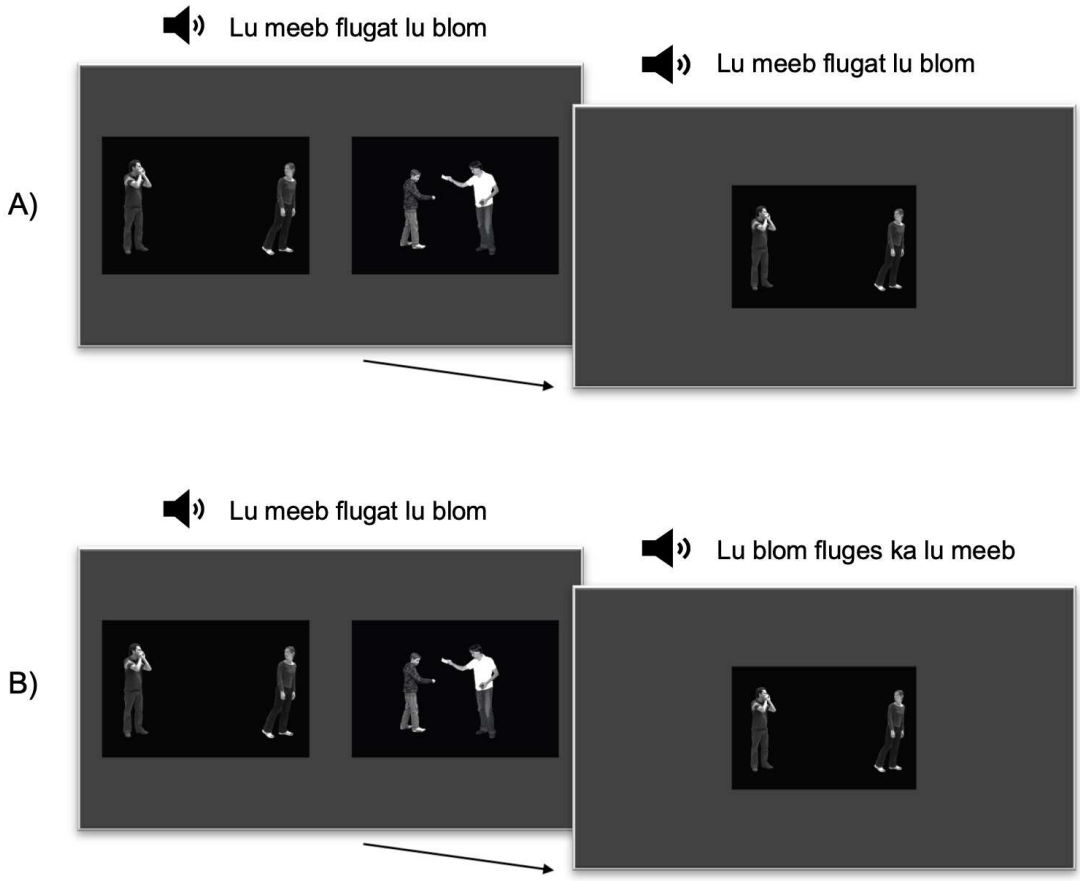


**Figure 3.** ***Learning trials with feedback, congruent (a) and incongruent (b)***

***Language Background and Debriefing Questionnaires***

At the end of Day 3, participants filled in a language background and debriefing questionnaire. The anonymised survey data can be downloaded from https://doi.org/10.17605/OSF.IO/NKSU8 and https://www.iris-database.org/ The first part of the questionnaire included questions on the participants' educational and language background, including the amount of formal grammar instruction received in the L1 and in any foreign languages spoken. The second part included specific questions on the experiment itself, aimed at probing participants' awareness of the structures and of the functional distinction between them ('Did you notice that a new

type of sentence was introduced on Day 2 (yesterday's session)?', and if Yes, 'What were the two types of sentence you learned, and what do you think the difference was between them?').

**Table 2.** ***Types of sentences included in the Grammaticality Judgment Task***

| Sentence type | Verb inflection | Example |
|---|---|---|
| Grammatical | Active | Lu meeb flug**at** lu blom |
| Grammatical | Passive | Lu blom flug**es ka** lu meeb |
| Ungrammatical (-at + ka) | Active | Lu meeb flug**at ka** lu blom |
| Ungrammatical (-es + Ø) | Passive | Lu blom flug**es** lu meeb |

### *Individual Difference Measures Taken on Day 3*

#### Serial Reaction Task.

A Serial Reaction Task (SRT) was administered to measure procedural learning abilities, following the paradigm used by Walker et al. (2020) and Lum, Gelgic, & Conti-Ramsden (2010). Participants saw a white square appear on one of four possible positions on screen (top, bottom, left and right), and had to press the corresponding arrow on their keyboard in response, as quickly and accurately as they could. The positions in which the square appeared followed a set sequence (bottom, top, right, left, right, top, bottom, right, top, left), which was repeated twice per block over five blocks (100 trials in total). The last block (Block 6) followed a different, pseudorandom sequence, repeated twice. Following Lum et al. (2010), the likelihood of the square appearing in any one particular position over the course of the pseudorandom sequence and the transitional probabilities between positions were kept consistent with those of the training sequence. To score the tasks, we followed Walker et al. (2020), subtracting the median RT for Block 5 from that for Block 6 (violation block).

#### LLAMA B3.

A vocabulary learning task (LLAMA B3) was included to measure verbal declarative memory. We replicated the design of the LLAMA B3 task (Meara & Rogers, 2019) using JavaScript library jsPsych (De Leeuw, 2015). Participants saw 20 drawings of novel fictional entities arranged in a grid on screen. Hovering over a drawing with the mouse cursor revealed a label with the written name of that entity. Participants were given 2 minutes to learn the names. At the end of the study period, the drawings appeared again, arranged in a different sequence. Participants were then given the names and asked to click on the relevant drawing (e.g., 'Click on the *taa*. If you are not sure, just guess'). All the drawings remained unchanged on screen throughout the test phase and could be selected at any time, and participants received no feedback on their answers.

## Results

A total of 70 participants were included in the analysis (Table 3). Four participants were excluded for failing to listen to the items before giving their responses (the criterion response time for this exclusion decision was under 1s on at least six trials per block, in any given block). One participant was excluded due to suspect unfair means (such as taking notes, based on response times over 10s and 100% accuracy from Block 1 of the cross-situational learning task on Day 1). One participant was excluded for failing to finish the Day 2 task in one sitting.

**Table 3.** ***Descriptive statistics for analysed sample***

| | Sex | Age | LLAMA B3 | SRT |
|---|---|---|---|---|
| | F | Years | Score | RT (ms) |
| *Group* | n | *M (sd)* | *M (sd)* | *M (sd)* |
| Surprisal (n = 36) | 29 | 30.9 (7.64) | 6.77 (4.07) | 45.82 (43.08) |
| Control (n = 34) | 25 | 31.1 (7.63) | 6.28 (4.54) | 38.86 (33.06) |

### Cross-situational Learning Task

We analysed accuracy data as binary outcome (correct / incorrect) at the trial level. We used generalized linear mixed-effect models (GLMER) for binomial data, which we implemented in R version 4.0.3 (R Core Team, 2020) using the *lme4* package (Bates, Mächler, Bolker, & Walker, 2015). Following Barr, Levy, Scheepers, & Tily (2013) we used the maximal random structure supported by the model, in order to control for as much variance as possible. For each model, we first created a formula containing the maximal fixed effect structure and the maximal random effect structure (random intercepts by subject and item as well as random slopes for subjects and items by each of the fixed effect predictors, and their interactions). We used the package *buildmer* (Voeten, 2020) to automatically identify the maximal random structure that would allow the model to converge. We then used *buildmer* again on the resulting formula to do stepwise backwards model selection using likelihood-ratio tests, eliminating fixed effect predictors one by one (starting from higher-level interactions) and only retaining them if they significantly improved model fit. All models were checked for overdispersion and none of them showed signs of being overdispersed. We report the coefficients of the mixed-effect models converted to odds ratios (*OR*) to provide a measure of effect size, together with the statistical significance of the effects (*p* values). Full descriptive statistics for the cross-situational task on all three days can be found in Appendix S2. Final statistical models for all tests can be found in Appendix S5.

#### *Learning Blocks*

Learning trials were included in the cross-situational task on Day 1 (Figure 4) and on Day 2 (Figure 5). We analysed data from the learning trials on Day 1 (blocks 1 – 5) and Day 2 (blocks 2 – 3 and 5 – 8) in two separate models, entering Group and Block (centred) as fixed effects for each. There were no significant differences in

performance between groups on learning blocks on either Day 1 or Day 2. There was a significant effect of Block on both Day 1 (*OR* = 1.28, 95% CI [1.21, 1.36], $p < .001$) and Day 2 (*OR* = 1.12, 95% CI [1.06, 1.18], $p < .001$).

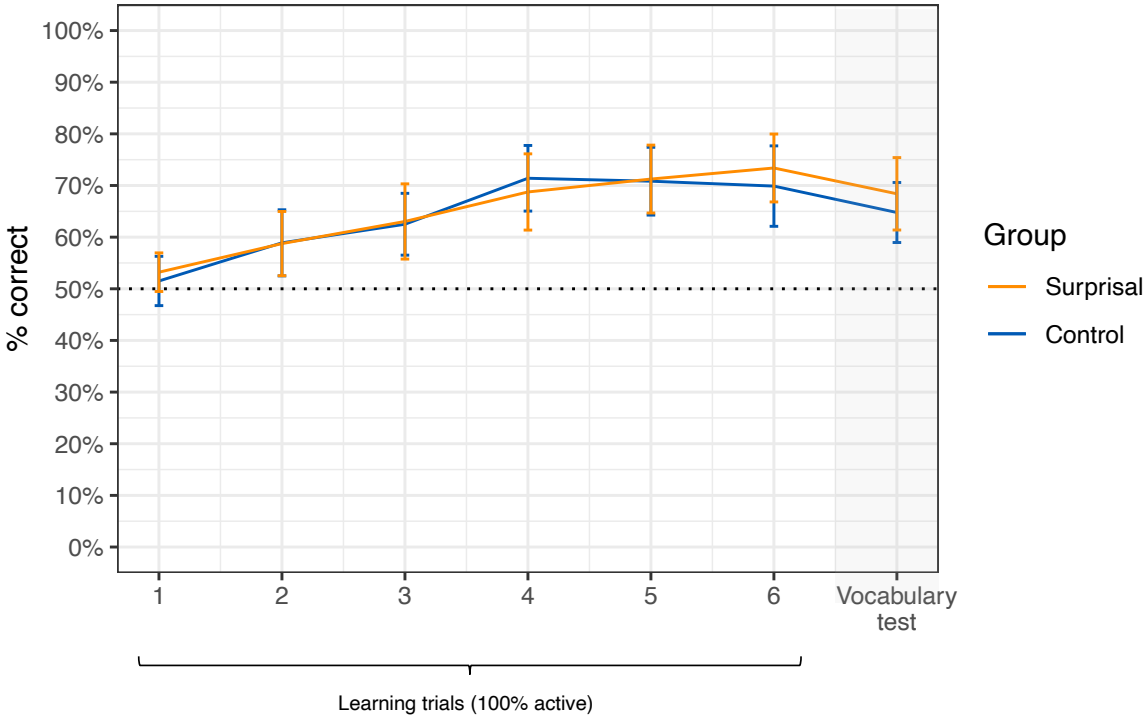


**Figure 4.** ***Accuracy on Day 1 by block. Error bars represent 95% CIs of group means (computed by averaging over subject means).***

### *Vocabulary Test Blocks*

Participants undertook five vocabulary test blocks over the course of the experiment: one on Day 1 (Test 1), three on Day 2 (Test 2, 3 and 4), and one on Day 3 (Test 5) (Figure 4, Figure 5 & Figure 6). We entered data from all these tests together into *glmer* models with Group and Test (centred) as fixed factors, creating two separate models for verbs and for nouns. For verbs, there was no significant effect of Group, only a main effect of Test (*OR* = 1.33, 95% CI [1.24, 1.41], $p < .001$). For nouns, there was a significant interaction between Test and Group (*OR* = 0.80, 95% CI [0.69, 0.93], $p = .004$). In post-hoc comparisons, a difference emerged in the transition from Test 3 to Test 4 (second and third test blocks on Day 2), which led to a significant improvement in accuracy for the Control group ($\chi^2(1) = 12.28$, $p = .009$) but not for the Surprisal group ($\chi^2(1) = 1.87$, $p = 1$). However, the difference in accuracy between the two groups was not significant at any point (Test 1: $\chi^2(1) = 0.13$, $p = 1$; Test 2: $\chi^2(1) = 0.27$, $p = 1$; Test 3: $\chi^2(1) = 0.005$, $p = 1$; Test 4: $\chi^2(1) = 1.51$, $p = 1$; $\chi^2(1) = 5.20$, $p = .113$).

### *Structure Test Trials after Feedback*

Individual structure test trials were inserted after feedback learning trials to test for any immediate (priming) effects as well as any cumulative effects of the experimental manipulation on structural knowledge. We analysed data from the Structure test

trials in a *glmer* model with Group and Trial number (scaled and centred) as fixed factors. We observed a main effect of Trial (*OR* = 1.40, 95% CI [1.17, 1.67], $p < .001$), but no effects of Group.

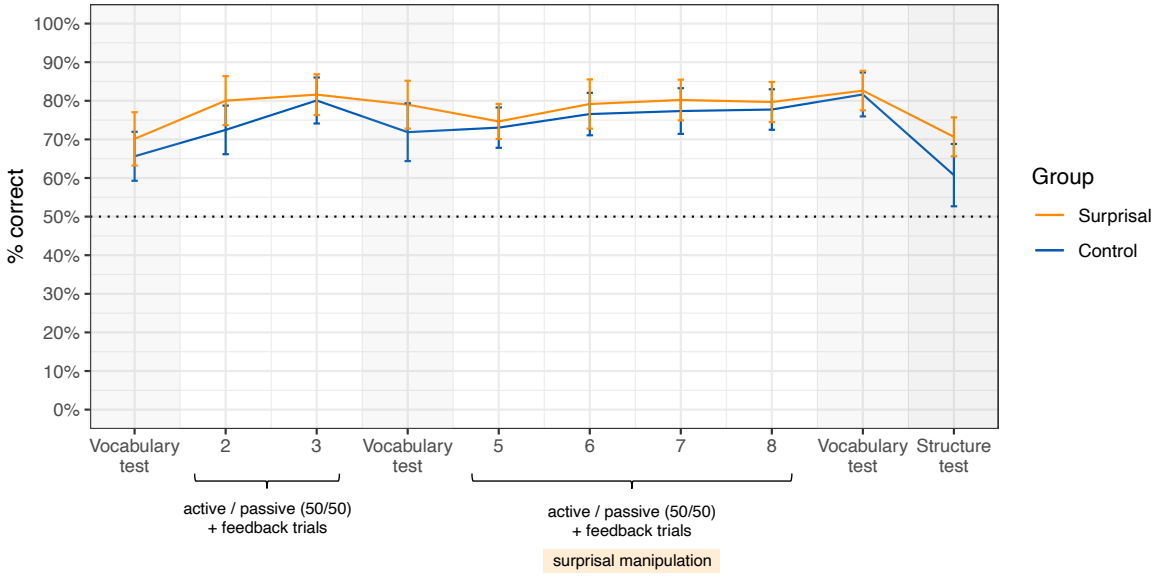


**Figure 5.** ***Accuracy on Day 2 by block. Error bars as in Figure 4.***

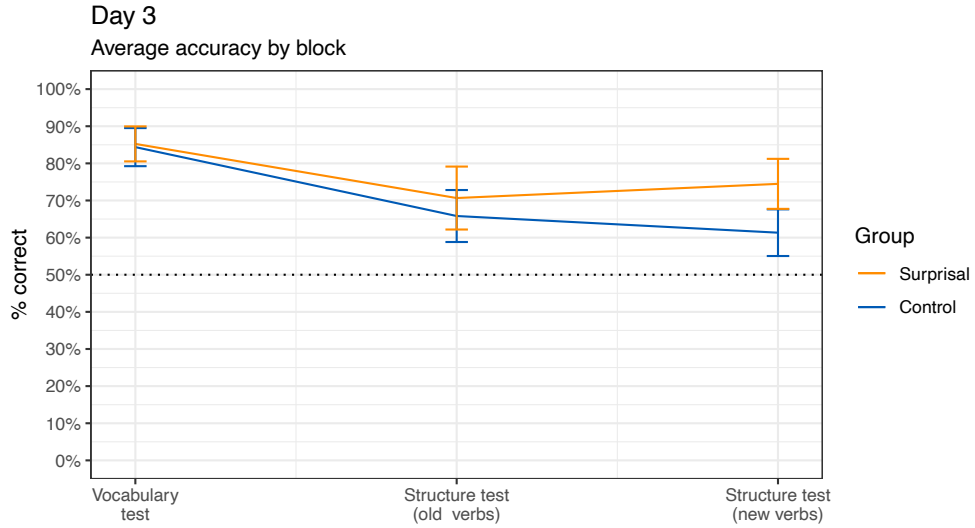


**Figure 6.** ***Accuracy on Day 3 by block. Error bars as in Figure 4.***

### *Structure Test Blocks*

We analysed data from each of the three structure test blocks in individual *glmer* models, entering Group, Structure (Active vs. Passive) and their interaction as predictors in the initial model for each.

*Day 2 (old verbs)*. In the structure test block on Day 2 (Figure 5), there was a numerical trend towards higher accuracy in the Surprisal group and for Active sentences, but no statistically significant effect of either Structure or Group. There was a great deal of variability between participants (see Appendix S4 for additional figures).
*Day 3 (old verbs)*. In this test block, we found a significant effect of Structure, with higher accuracy for Active relative to Passive (*OR* = 2.80, 95% CI [1.50, 5.23], *p* = .001), but no effect of Group.[2]

*Day 3 (new verbs)*. In the generalisation structure test block, there were significant main effects of Group (*OR* = 2.50, 95% CI [1.25, 4.99], *p* = .009) and Structure (*OR* = 2.82, 95% CI [1.42, 5.59], *p* = .003): Participants in the Surprisal group were more accurate than those in the Control group, and both groups had higher accuracy for active structures compared to passives (Figure 7).

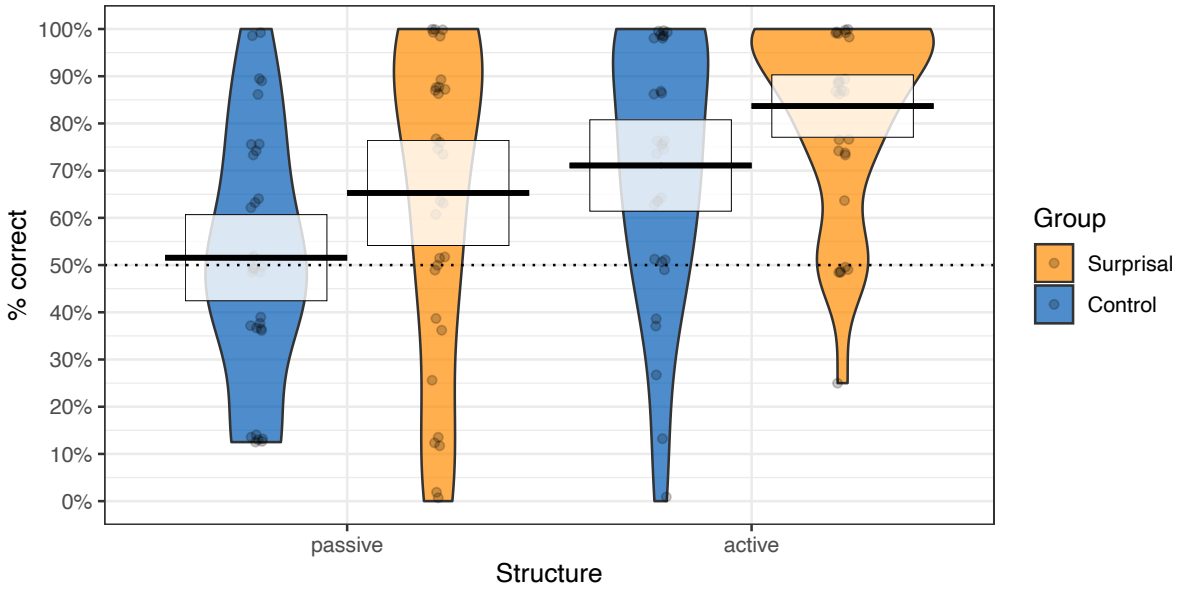


**Figure 7.** ***Accuracy on Day 3 structure test block (new verbs). Horizontal bars represent group means, shaded rectangles 95% CIs.***

## Grammaticality Judgment Task

Descriptive statistics for this task can be found in Appendix S2, and the full statistical models are reported in Appendix S5. We analysed both raw endorsement rates, to capture differences in endorsement bias between groups, and *d'* scores to obtain a measure of sensitivity to grammaticality in the two groups. Endorsement data was collected as a binary response (Yes / No) so we analysed it using a generalized linear mixed-effect model (GLMER) for binomial data, following the same procedure we use

[2] We can only draw limited conclusions from the results of the Day 3 (trained verbs) test block, however, as this block was affected by a counterbalancing error which meant that half of the participants (equally spread among groups) saw the exact same items as in the Day 2 structure test (while the other half saw the same pictures but described using the opposite structure, which was the intended design). This does not affect the following test block (Day 3, new verbs), which used entirely novel Agent – Verb – Patient combinations. See Appendix S4 for a detailed figure of the Day 3 (trained verbs) block, including accuracy by group and counterbalancing status.

to analyse accuracy data from the cross-situational learning task. Group, Verb type (active vs. passive) and Grammaticality were entered as fixed predictors in the model.

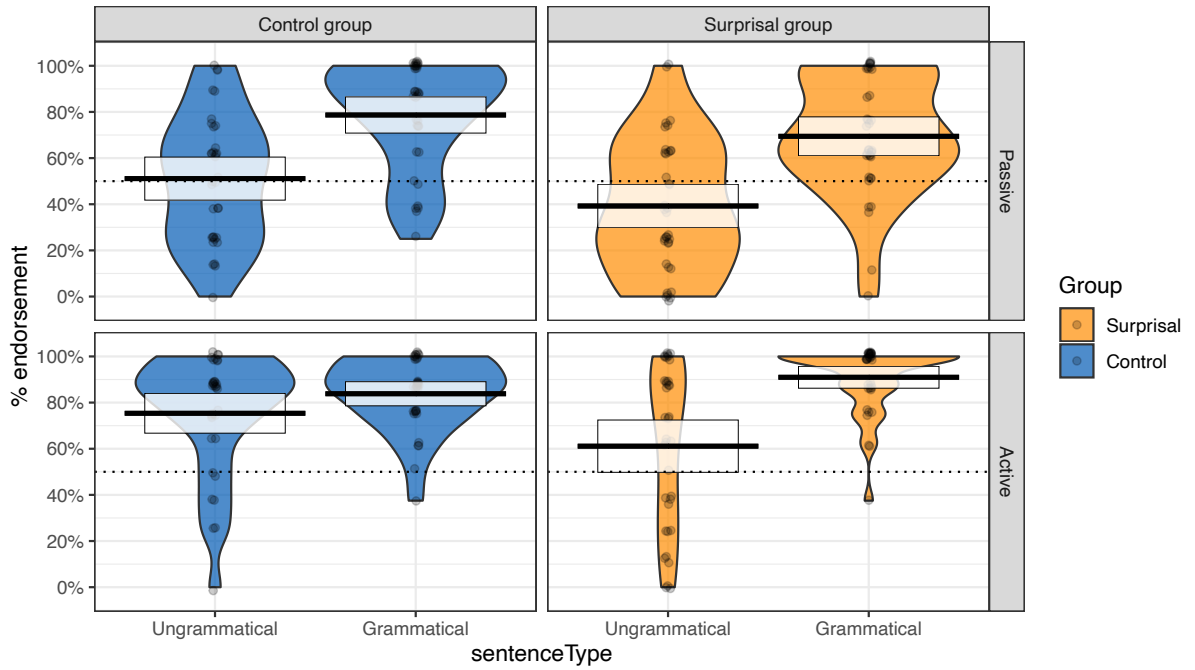


**Figure 8.** ***Endorsement rates in the Grammaticality Judgment Task (Day 3), by Group, sentence Grammaticality and Verb type. Horizontal bars represent group means, shaded rectangles 95% CIs.***

We found a three-way interaction between Group, Verb type and Grammaticality (*OR* = 4.44, 95% CI [1.85, 10.64], $p$ = .001) (Figure 8). Post-hoc comparisons showed that this was driven by the effect of Grammaticality varying across groups, specifically for Active sentences. Participants in the Surprisal group were significantly more likely to endorse grammatical relative to ungrammatical sentences, whether they contained Active ($\chi^2(1) = 29.22$, $p < .001$) or Passive verb forms ($\chi^2(1) = 21.86$, $p < .001$). The Control group, on the other hand, showed the same effect of grammaticality with Passive sentences ($\chi^2(1) = 19.54$, $p < .001$) but not with Active ones ($\chi^2(1) = 1.57$, $p$ = .84). The effect of grammaticality was of similar magnitude for Passive sentences in both groups (SG: *OR* = 0.18, 95%CI [0.06, 0.59], $p < .00$; CG: *OR* = 0.19, 95%CI [0.05, 0.66], $p < .001$) and in Active sentences for the Surprisal group (*OR* = 0.12, 95%CI [0.02, 0.62], $p$ = .001). The Control group were more likely than the Surprisal group to endorse sentences with a Passive verb in general, regardless of their grammaticality ($\chi^2(1) = 5.74$, $p$ = .033). In general, endorsement across groups was higher for Active then for Passive sentences, both grammatical (($\chi^2(1) = 10.30$, $p$ = .003) and ungrammatical ($\chi^2(1) = 29.66$, $p < .001$).

To estimate participants' ability to discriminate grammatical from ungrammatical sentences, regardless of endorsement bias, we calculated $d'$ scores for different item types and entered them in a mixed ANOVA using package *ez* (Lawrence, 2016), with Group and Verb type as predictors. The ANOVA returned a significant interaction between Group and Verb type ($F(1,68) = 4.590$, $p$ = .036) but no significant main effects

of either Group ($F(1,68) = 1.809$, $p = .183$) or Verb type ($F(1,68) = 0.230$, $p = .633$). We carried out post-hoc comparisons with the Bonferroni correction using package *emmeans* (Lenth et al., 2021), which showed a significant difference in $d'$ scores between groups for Active sentences ($F(1,68) = 5.505$, $p = .04$) but not for Passive ones ($F(1.68) = 0.028$, $p = 1$).

### Cognitive Measures

Basic descriptive statistics for LLAMA B3 and SRT scores can be found in Table 3; see Appendix S1 for detailed statistics. The groups did not significantly differ in their scores, and the effects of Group we reported for the cross-situational learning task and grammaticality judgment task were not affected by the inclusion of these cognitive measures in the analysis. See Appendix S1 for a detailed report of the analyses that included these measures as variables.

### Debriefing Questionnaire (RQ2)

21 out of 36 subjects in the Surprisal group and 14 out of 34 subjects in the Control group developed sufficient explicit knowledge of the structures to be able to verbalise their respective functions. To assess whether the experimental manipulation had made participants in the Surprisal group more likely to develop explicit knowledge of the Active / Passive distinction, we constructed a simple logistic regression with explicit knowledge as a binary outcome and Group as predictor. While the Surprisal group had a numerically higher rate of explicit knowledge, the effect was not significant ($OR = 0.50$, 95% CI [0.19, 1.28], $p = .15$).

## Discussion

Our first research question concerned the effect of surprisal on structural knowledge. We hypothesised that surprisal at the item level would lead to stronger abstract structural knowledge of the passive structure in the Surprisal group: Our results partially supported this hypothesis. We found that participants in the Surprisal group performed significantly better than those in the Control group in both a structure comprehension test and a grammaticality judgment task. Crucially, the structure test used novel verbs, which shows that the Surprisal group had developed stronger abstract knowledge than the Control group, and were able to use that knowledge to generalise structure to a new lexicon. However, we did not observe an effect of Group on the other structure test blocks or structure test trials in earlier blocks, which all used familiar verbs. Additionally, the effects we observed, were not limited to the passive construction as we had hypothesised, given that the manipulation was only on passive items. In the comprehension test, the advantage for the Surprisal group was found across both structures. In the Grammaticality Judgment Task, against our expectations, the main difference between groups emerged on active sentences, where only the Surprisal group showed a significant ability to distinguish grammatical from ungrammatical sentences.

Our secondary hypothesis was that surprisal would also lead to greater awareness of the functional distinction between active and passive constructions, measured as the ability to verbalise the distinction in retrospective verbal report. While there was a

numerical advantage for the Surprisal group, this was not statistically significant. We consider these findings below, offering possible interpretations for the observed pattern of results and discussing the limitations of the current study.

**Structural Accuracy in Comprehension**

Against our expectations, we did not find an effect of Group in the structure tests which used familiar verbs (on either the structure tests blocks or the structure tests trials following feedback trials). We discuss potential explanations for these findings in the section on 'Study limitations' below. However, in the structure test on Day 3 (new verbs), we found a main effect of Structure, and one of Group: Both groups were better at selecting the correct interpretation of active sentences than they were for passive ones, and the Surprisal group was overall more accurate than the Control group. The effect of structure is compatible with our experimental design: Given that participants had received more and earlier exposure to this structure than to the passive, it is not surprising that they developed higher accuracy on it. We also expected the Surprisal group to perform better than the Control group in the structure test, which was confirmed. However, the effect was found for both Active and Passive structures (and was numerically greater for active ones), whereas we had expected to find an advantage specifically for passive sentences, given that they were the target of our experimental manipulation.

One possible explanation is that the mere presence of surprising trials led to greater attention and therefore better overall learning in the Surprisal group. In a series of cross-situational learning studies of vocabulary learning, Fitneva & Christiansen (2011, 2017) found that experiencing error (i.e., initially forming incorrect label-referent mappings) led to better learning in adults. Crucially, this effect was not limited to the words that participants had initially assigned to the wrong referent, but to the whole set of items, suggesting that experiencing error may have led to greater attention and better encoding of information overall (Fitneva & Christiansen, 2017).

A second possibility is that the effect was due to an interplay between the two structures: better knowledge of the passive construction could have led to higher accuracy on active trials, by providing negative evidence that helped participants rule out the incorrect alternative. In our structure test, the competitor (incorrect) picture always depicted the same action happening with Agent and Patient roles reversed, meaning the two constructions were effectively put in competition against each other. If the sentence was in the active form, e.g., *Lu meeb flug**at** lu prag* ('The girl calls the boy'), then the target picture would depict a girl calling a boy, while the competitor would depict a girl being called by a boy. This means that a sentence with the same nouns in the same positions as the target sentence could be used to describe the competitor picture, but only if it had different morphosyntax, that is, *Lu meeb flug**es ka** lu prag,* ('The girl is called by the boy'). Being sensitive to this distinction would help participants make the correct choice by ruling out the competitor picture, that is, by providing negative evidence of what the active sentence could *not* describe. Crucially, however, this requires specific sensitivity to the morphosyntactic distinction, which would in turn depend on accurate knowledge of the passive construction, as well as the active. Relying only on vocabulary would not be of help

in this context, as both pictures could be described by sentences containing the same verb and nouns in the same order.

Yet another potential explanation for our findings is that the surprisal feedback trials did lead to better structural learning, but not in the way we had hypothesised. It is possible that what drove the effect of the surprisal feedback trials was actually the juxtaposition of an active and passive sentence used in sequence to describe the same event, rather than the passive feedback sentence being better encoded due to it being unexpected. This would have showed learners that the two structures could be used to describe the same event, potentially prompting them to pay more attention to the specific form-meaning mappings in the two structures. If learners follow a 'uniqueness principle' and assume that any given meaning can only be encoded by one grammatical form (Pinker, 2009), then the presence of two superficially equivalent forms may trigger a search for functional distinctions that may justify the existence of both forms in the grammar. We have no way to confirm or rule out this explanation given the currently available data. One future development of this research, however, will be to include a measure of item memory, testing for specific memory of the feedback sentences received in the critical feedback trials. If participants do show better memory for passive feedback sentences encountered in the surprising condition, this will lend support to our original hypothesis, that the surprisal manipulation improved memory for specific, individual items, which in turn lead to better generalisation. However, this would not entirely rule out a role for the second potential mechanism just described (i.e., juxtaposition of two structures leading to more accurate representations of structure-meaning mappings). In order to fully investigate this point, further research could include a different way to generate surprisal, that does not result in juxtaposition of an active with a passive sentence describing the same picture. If the same effects are observed, it would suggest that the effect of our experimental manipulation was not primarily driven by an artefact of our experimental design (the juxtaposition of two structures for the same event) but, rather, but the surprisal phenomenon itself.

In sum, further research needs to attempt to identify the explanatory power of these two accounts. However, it might also be worth bearing in mind that these two mechanisms could in fact—at least some of the time—be two sides of the same coin, working reciprocally, in tandem; that is, surprisal may serve to highlight meaning-(or function-) bearing linguistic contrasts, and, in turn, meaning-bearing contrasts may be a cause of surprisal events.

**Development of Explicit Knowledge**

Our second experimental hypothesis was that the surprisal manipulation may lead participants to develop a higher degree of awareness of the functional distinction between active and passive sentences. The data we collected does not allow us to satisfactorily answer this question, unfortunately. The debriefing questionnaire we used to measure explicit rule knowledge showed a numerical difference between groups, with higher rates of awareness among the Surprisal group; however, this was not significant in a statistical test. It is possible that the questionnaire may simply have been underpowered: due to the lack of previous research using a similar manipulation, we did not have a reference effect size which we could expect to see in

this study, with regards to explicit knowledge[3]. Future developments of this research could employ larger sample sizes to address the possibility that our manipulation did have an effect on rule awareness, but that the effect was too small to be detected with sufficient confidence in our sample.

In addition to the potential lack of power, our measure of awareness was admittedly not a fine-grained one. It merely set a threshold based on retrospective verbal report, to divide participants into two categories (aware and unaware). Retrospective verbal report has been criticised for a potential lack of sensitivity to awareness; for instance, lack of confidence may lead to underreporting in some participants (Rebuschat, 2013). It could still be that awareness itself was emerging in a graded manner as structural representations were becoming stronger and more stable in the Surprisal group, in a way that is not captured by our cut-off point (the ability to verbalise the functional distinction between structures). This kind of graded emergence of awareness was the other potential mechanism that we hypothesised may have led to increased awareness in the Surprisal group, compatible with the radical plasticity theory of the relation between implicit and explicit knowledge (Cleeremans, 2008, 2011).

A related question, answers to which can only remain speculative for now, concerns the extent to which accuracy in the structure test may have been driven by explicit knowledge. While the difference between groups in terms of their reported awareness of the structure was not statistically significant, there was a numerical advantage for the Surprisal group. A higher degree of structural awareness may have helped participants in the Surprisal group to perform better in the structure test, once they did become aware (or were 'on their way to' awareness). However, given that awareness was assessed at the end of the study by verbal report, we do not know at which point participants did become sufficiently aware of the distinction to influence accuracy. Knowing that tipping point would be a prerequisite for any analysis aiming to use awareness as a predictor for accuracy. To address both of these points—the gradual emergence of awareness, and the extent to which it contributed to performance in the structure test—future research would need to include more fine-grained measures of awareness administered as the trials progressed, such as source attribution (Dienes & Scott, 2005) or the use of multiple direct and indirect tests to tease apart the contribution of different types of knowledge (Ellis Rod, 2009). The challenge for that line of research is to avoid 'reactivity', whereby the probe of awareness itself promotes, or interferes with, actual awareness (Bowles, 2010).

**Grammaticality Judgment Task**

In the grammaticality judgment task, we found further evidence that the Surprisal group had developed better structural knowledge than the Control group, broadly

[3] The effect size we observed in the Day 3 Structure Test on new verbs was Cohen's $d$ = .64 (a medium effect size according to Cohen's (1988) benchmark, but a small one in the context of L2 acquisition research (Plonsky & Oswald, 2014)). Based on this effect size, we carried out a post-hoc power analysis in G*Power, which showed that the study had .84 power to detect the effect which we observed in structural comprehension. However, this does not provide an indication of the power the study had to detect a potential effect in awareness, which may be smaller than the one on structural comprehension (see section on Study limitations).

supporting our experimental hypothesis. However, we did not find the exact effect that we had anticipated, that is, greater accuracy in discriminating grammatical from ungrammatical sentences by the Surprisal group on passive items, compared to the Control group. Instead, *both* groups were significantly more likely to endorse grammatical sentences relative to ungrammatical ones in the passive structure, and the magnitude of the effect was the same in both groups. The two groups, however, differed in their overall likelihood to endorse passive sentences—irrespective of grammaticality—in that endorsement of passives was lower in the Surprisal group. This was not predicted by our experimental hypothesis. It may reflect a greater sense in the Surprisal group of the fact that the passive was an entirely new structure, while the Control group was more accepting of all sentences that resembled items they encountered during training.

On the other hand, a significant interaction between grammaticality and group—indicating that the groups differed in their ability to discriminate between grammatical and ungrammatical sentences—did emerge, but only for Active items. Here, the difference was remarkable: The Surprisal group showed a difference in endorsement rates between grammatical and ungrammatical sentences which was statistically significant and comparable in size to the effect observed for passive items. The Control group, by contrast, showed practically no difference in their endorsement of grammatical and ungrammatical items, and were equally likely to endorse any sentence containing an active verbal form (ending in *-es*), regardless of whether it was used in a grammatical way. An analysis of *d'* scores confirmed that the Surprisal and Control group differed significantly in their ability to discriminate grammatical from ungrammatical sentences, but only when they were in the active form.

This pattern may be explained by considering the way in which ungrammatical items were constructed in the grammaticality judgment task. These items mixed morphosyntax from different structures to create sentences that were unattested in the input participants had thus far received. Specifically, ungrammatical active sentences contained the active verbal suffix *-at* followed by the passive agent marker *ka*, while ungrammatical passive sentences contained the passive verbal suffix *-es* without the agent marker *ka* (see Table 2 for example stimuli). It appears that participants in the Control group were equally likely to endorse any sentence that contained chunks they had already encountered in training: either a verb with an active suffix, or a verb with a passive suffix followed by *ka*. When one of these chunks was broken—as happened in the case of ungrammatical passive sentences, which had a passive verb suffix but no *ka*—they were sensitive to this violation, resulting in lower endorsement rates. However, when a chunk was found in its entirety, as previously attested (verb + active suffix), but followed by a novel element—as in ungrammatical active sentences, where the active verb inflection was followed by *ka*—they did not perceive this as violating an established pattern. This could indicate that they were paying less attention to the material that followed the verbal inflection in the sentence, compared to the Surprisal group.

By contrast, the Surprisal groups showed equal sensitivity to ungrammatical usage of both active and passive verb forms, showing that they also paid attention to the material following the verbal inflection, resulting in lower endorsement for active

inflections being followed by a novel item (the *ka* marker). This suggests that they had developed a more sophisticated kind of knowledge than the Control group. They had not only acquired the individual forms for active and passive inflection (and its associated marker), but, crucially, they had learned better that the two forms were associated with a different order of Agent and Patient. This suggests that they paid attention to both the material that followed the verb as well as that preceding it. In the grammaticality judgment task, this allowed them to discriminate between grammatical and ungrammatical usage of both verbal suffixes.

The lack of sensitivity shown by the Control group to grammaticality in active items is seemingly at odds with the results of the structure test, where they were able to pick the correct interpretation for active sentences with reasonably good accuracy. However, in the active items in the structure test, participants did not technically need to pay attention to the noun following the verb (the Patient) to answer correctly. Just correctly identifying the first noun as the Agent of the action, in combination with the active inflection, would suffice to answer correctly. Therefore, considering the results of both the structure and the grammaticality judgment task together, it is possible that participants in the Control group had settled on a basic heuristic, namely identifying the first noun as the Agent of the sentence (independently of verbal inflection), which was sufficient to answer correctly to active sentences in the structure test. This is also compatible with the fact that they were essentially at chance level in their responses to passive items in the structure test. By contrast, the Surprisal group could rely on additional cues for determining the correct meaning of the sentences (by attending to the material that came after first noun), resulting in higher accuracy on *both* active and passive items.

Crucially, however, it is not the case that Control participants simply never paid attention to the second noun in sentences. There was a specific task in the study—namely, the vocabulary test blocks—which could only be performed correctly by paying attention to the Patient noun (always in second position, since all sentences in vocabulary test blocks were active). There was no significant difference between groups in vocabulary test blocks, indicating that both groups were attending to the relevant noun. When the task did not specifically demand it (as in the structure test), however, the Control group did not seem to attend to the material following the verbal inflection. This suggests that they had developed little sensitivity to the relation between noun position, verb form, and sentence meaning. The Control group learned that different verbal forms existed, but their knowledge of the structures they were found in, with the relevant form-meaning connections—that is, the different assignment of Patient and Agent roles—was reduced, relative to that of the Surprisal group. In turn, this resulted in lower accuracy in the Control group in the comprehension task, too, for passive sentences.

**Study Limitations**

Against our expectations, we did not observe an effect of group in structure tests that used previously trained verbs, in either the structure test blocks (on Days 2 and 3) or in the individual test trials following feedback (on Day 2). To some extent, these findings may be explained by limitations in the structure tests themselves. As we previously mentioned, the first structure test on Day 3 (old verbs) was affected by a

counterbalancing problem. The lack of an effect on the individual test trials following feedback, too, could be due to limitations in the study setup. These structure test trials were placed after 'critical' learning trials, that is, those which included incongruent passive feedback in the Surprisal group, and their congruent passive counterparts in the Control group. We hypothesised that surprisal may lead to a stronger structural priming effect in the Surprisal group, which may manifest itself as higher accuracy on passive structure test trials. However, congruent trials in the Control group involved passive feedback presented after a passive sentence, meaning that Control participants were exposed to two passive sentences in a row, leading to potential cumulative priming. Therefore, it is possible that even if any effect due to surprisal was present, its effects relative to the Control group may have been obscured by cumulative priming effects in the Control group. This could have potentially cancelled out any differences between groups.

However, these problems do not affect the first structure test block using old verbs, which was administered at the end of Day 2. Why did we observe an effect of Group in the generalisation test on Day 3, but not in the old verbs test on Day 2? One possibility is that the surprisal effect, which we hypothesised to affect memory formation for critical passive sentences, may have required overnight sleep for memory consolidation and abstraction to take place[4]. The possible need for overnight consolidation was one of the reasons behind the decision to add tests on Day 3, in addition to the test at the end of Day 2. Under this interpretation, we should have observed an effect on first Day 3 structure test (old verbs), too. However, the technical error affecting this test blocks means that we have no conclusive evidence on this point. Further research replicating this design would be needed to provide evidence in support of this hypothesis about the role of sleep consolidation.

There are, however, other indicators from the study suggesting that the surprisal manipulation did not work as intended, i.e., by generating stronger memories for passive sentences when presented in surprising feedback. One point, which we already raised in the discussion, was that the effect on structural comprehension was found for both structures, not just the passive. Since only the passive was meant to be affected by our surprisal manipulation, it seems that the manipulation did not have the effect it was meant to have. We mentioned in the discussion the possibility that juxtaposition between structures in incongruent trials may have caused the effect we observed, by leading to higher awareness of the rule. The results of the debriefing questionnaire are not conclusive in this respect: they show a numerical difference between groups, which, however, is not significant. We have discussed the possibility that, while the study appeared sufficiently powered to detect the effect on structural comprehension, it may not have been sufficient to detect potentially smaller effects on awareness (footnote on p. 25). Indeed, sensitivity to differences between structures as a result of juxtaposition could have been stronger in the Surprisal group than in the Control group, leading to better performance in comprehension, but still not strong enough to lead to the level of explicit awareness needed to verbalise the distinction.

[4] While we are not aware of any research on the consolidation of syntactic structure, work using novel (artificial) L2 morphology shows an effect of overnight consolidation on the acquisition of new systematic patterns (Mirković et al., 2019; Tamminen et al., 2015).

Similarly, in the Grammaticality Judgment Task we observed an effect of Group but only for active sentences, not for passive ones, which is at odds with the fact that our manipulation was intended to target passive sentences. In the discussion, we offered a potential explanation of the results of the Grammaticality Judgment Task based on different patterns of attention in the two groups: we hypothesised that the Surprisal groups had developed stronger sensitivity to the fact that different morphosyntax on the verb correlated with different orderings of Agent and Patient, while the Control group relied on an 'Agent first' heuristic which accounted for their low performance in structural comprehension of passive sentences. However, the juxtaposition of active and passive sentences in the feedback, a potential limitation (confound) in the design, could plausibly have caused such an effect, too.

Finally, we should point out that the manipulation we used, whatever its effects, was quite subtle: there were only 4 critical trials per block, for a total of 16 over the whole experiment. Additionally, the expectation for congruent feedback—which was necessary for participants to experience surprisal at incongruent feedback—was only set up over the course of the first two learning blocks on Day 2 (a total of 24 trials with congruent feedback), which may have been insufficient to set up sufficiently strong expectations for congruent feedback to influence some of the dependent variables examined. In sum, some of the limitations and incongruities in this study may also be the result of a relatively weak manipulation. Future developments of this study could use a stronger surprisal manipulation, which may shed more light on some of the issues raised in this section.

## Conclusion

In this study, we examined the effect of expectation violation on the acquisition of novel syntactic structures. Specifically, we examined the acquisition a minority syntactic structure (passive) introduced after the default structure (active) had been consolidated. We hypothesised that presenting instances of the passive structure in a way that violated expectations (surprisal) would lead to better acquisition of the passive structure itself, and greater awareness of its function. Our predictions with regards to accuracy were mainly supported: Although the pattern of results did not support the prediction of an isolated effect on only the passive structures, it clearly demonstrated that the Surprisal group developed stronger and more accurate structural representations than the Control group, for both constructions. In contrast, the experimental manipulation did not lead to statistically significantly sufficient levels of awareness to lead to knowledge that could be articulated explicitly, despite a numerical trend in that direction. The lack of statistical significance could be due to a number of design and methodological limitations, however, and the role of explicit knowledge should be investigated further. Nevertheless, it seems intriguing to us that a very simple manipulation, on a relatively small number of trials, had quite significant consequences for the representations developed by the two groups, and seemed to lead to different patterns of attention, too. Further research will be needed to investigate the effects we found, and to pinpoint their exact origin, among the different explanations we offered.

## References

Arai, M., van Gompel, R. P. G., & Scheepers, C. (2007). Priming ditransitive structures in comprehension. *Cognitive Psychology, 54*(3), 218–250. doi:10.1016/j.cogpsych.2006.07.001

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013-4). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3). doi:10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48. doi:10.18637/jss.v067.i01

Bock, K. (1986). Syntactic persistence in language production. *Cognitive Psychology, 18*(3), 355–387. Retrieved from http://www.linguisticsnetwork.com/wp-content/uploads/Syntactic-Persistance_Bock-1986-.compressed.pdf

Bock, K., Dell, G. S., Chang, F., & Onishi, K. H. (2007). Persistent structural priming from language comprehension to language production. *Cognition, 104*(3), 437–458. doi:10.1016/j.cognition.2006.07.003

Bock, K., & Griffin, Z. M. (2000). The persistence of structural priming: transient activation or implicit learning? *Journal of Experimental Psychology. General, 129*(2), 177–192. doi:10.1037//0096-3445.129.2.177

Bowles, M. A. (2010). *The think-aloud controversy in second language research* [X, 172 p. : ill. ; 24 cm.]. New York ; Abingdon: Routledge.

Brod, G., Hasselhorn, M., & Bunge, S. A. (2018). When generating a prediction boosts learning: The element of surprise. *Learning and Instruction, 55*, 22–31. doi:10.1016/j.learninstruc.2018.01.013

Bybee, J. L., & Hopper, P. J. (2001). *Frequency and the Emergence of Linguistic Structure*. Amsterdam: John Benjamins Publishing.

Cerezo, L., Caras, A., & Leow, R. P. (2016). The effectiveness of guided induction versus deductive instruction on the development of complex Spanish gustar structures: An analysis of learning outcomes and processes. *Studies in Second Language Acquisition, 38*(2), 265–291. doi:10.1017/S0272263116000139

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review, 113*(2), 234–272. doi:10.1037/0033-295X.113.2.234

Chang, F., Janciauskas, M., & Fitz, H. (2012). Language adaptation and learning: Getting explicit about implicit learning. *Language and Linguistics Compass, 6*(5), 259–278. doi:10.1002/lnc3.337

Cintrón-Valentín, M. C., & Ellis, N. C. (2016). Salience in Second Language Acquisition: Physical Form, Learner Attention, and Instructional Focus. *Frontiers in Psychology, 7,* 1284. doi:10.3389/fpsyg.2016.01284

Cleeremans, A. (2008). Consciousness: the radical plasticity thesis. *Progress in Brain Research, 168,* 19–33. doi:10.1016/S0079-6123(07)68003-0

Cleeremans, A. (2011). The Radical Plasticity Thesis: How the Brain Learns to be Conscious. *Frontiers in Psychology, 2,* 86. doi:10.3389/fpsyg.2011.00086

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods, 47*(1), 1–12. doi:10.3758/s13428-014-0458-y

De Loof, E., Ergo, K., Naert, L., Janssens, C., Talsma, D., Van Opstal, F., & Verguts, T. (2018). Signed reward prediction errors drive declarative learning. *PloS One, 13*(1), e0189212. doi:10.1371/journal.pone.0189212

Dienes, Z., & Scott, R. (2005). Measuring unconscious knowledge: distinguishing structural knowledge and judgment knowledge. *Psychological Research, 69*(5), 338–351. doi:10.1007/s00426-004-0208-3

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24*(2), 143–188.

Ellis, N. C. (2016). SALIENCE, COGNITION, LANGUAGE COMPLEXITY, AND COMPLEX ADAPTIVE SYSTEMS. *Studies in Second Language Acquisition, 38*(2), 341–351. doi:10.1017/S027226311600005X

Ellis, N. C. (2017). Salience in usage-based SLA. In S. M. Gass, P. Spinner, & J. Behney (Eds.), *Salience in Second Language Acquisition* (1st ed., pp. 21–40). doi:10.4324/9781315399027

Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). *Usage-based approaches to language acquisition and processing: cognitive and corpus investigations of construction grammar*. Hoboken, NJ: Wiley.

Ellis Rod. (2009). Implicit and explicit knowledge in second language learning, testing and teaching. In R. Ellis, S. Loewen, C. Elder, R. Erlam, J. Philp, & H. Reinders (Eds.), *Implicit and explicit knowledge in second language learning, testing and teaching* (Vol. 42).

Farmer, T., Fine, A., Yan, S., Cheimariou, S., & Jaeger, F. (2014). Error-Driven Adaptation of Higher-Level Expectations During Reading. *Proceedings of the Annual Meeting of the Cognitive Science Society, 36*(36). Retrieved from https://escholarship.org/uc/item/00t2m3wr

Fazekas, J., Jessop, A., Pine, J., & Rowland, C. (2020). Do children learn from their prediction mistakes? A registered report evaluating error-based theories of language acquisition. *Royal Society Open Science, 7*(11), 180877. doi:10.1098/rsos.180877

Fazio, L. K., & Marsh, E. J. (2009). Surprising feedback improves later memory. *Psychonomic Bulletin & Review, 16*(1), 88–92. doi:10.3758/PBR.16.1.88

Ferreira, V. S., & Bock, K. (2006). The functions of structural priming. *Language and Cognitive Processes, 21*(7–8), 1011–1029. doi:10.1080/016909600824609

Fine, A. B., & Jaeger, T. F. (2016). The role of verb repetition in cumulative structural priming in comprehension. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 42*(9), 1362–1376. doi:10.1037/xlm0000236

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid Expectation Adaptation during Syntactic Comprehension. *PloS One, 8*(10), e77661. doi:10.1371/journal.pone.0077661

Fitneva, S. A., & Christiansen, M. H. (2011). Looking in the wrong direction correlates with more accurate word learning. *Cognitive Science, 35*(2), 367–380. doi:10.1111/j.1551-6709.2010.01156.x

Fitneva, S. A., & Christiansen, M. H. (2017). Developmental Changes in Cross-Situational Word Learning: The Inverse Effect of Initial Accuracy. *Cognitive Science, 41 Suppl 1*, 141–161. doi:10.1111/cogs.12322

Greve, A., Cooper, E., Kaula, A., Anderson, M. C., & Henson, R. (2017). Does prediction error drive one-shot declarative learning? *Journal of Memory and Language, 94*, 149–165. doi:10.1016/j.jml.2016.11.001

Greve, A., Cooper, E., Tibon, R., & Henson, R. N. (2019). Knowledge is power: Prior knowledge aids memory for both congruent and incongruent events, but in different ways. *Journal of Experimental Psychology. General, 148*(2), 325–341. doi:10.1037/xge0000498

Hartsuiker, R. J., Kolk, H. H. J., & Huiskamp, P. (1999). Priming Word Order in Sentence Production. *The Quarterly Journal of Experimental Psychology Section A, 52*(1), 129–147. doi:10.1080/713755798

Hartsuiker, R. J., & Westenberg, C. (2000). Word order priming in written and spoken sentence production. *Cognition, 75*(2), B27-39. doi:10.1016/s0010-0277(99)00080-3

Jackson, C. N. (2018). Second language structural priming: A critical review and directions for future research. *Second Language Research, 34*(4), 539–552. doi:10.1177/0267658317746207

Jackson, C. N., & Ruf, H. T. (2017). The priming of word order in second language German. *Applied Psycholinguistics, 38*(2), 315–345. doi:10.1017/S0142716416000205

Jaeger, T. F., & Snider, N. E. (2013). Alignment as a consequence of expectation adaptation: syntactic priming is affected by the prime's prediction error given both prior and recent experience. *Cognition, 127*(1), 57–83. doi:10.1016/j.cognition.2012.10.013

Kaan, E., & Chun, E. (2018a). Priming and adaptation in native speakers and second-language learners. *Bilingualism: Language and Cognition, 21*(2), 228–242. doi:10.1017/S1366728916001231

Kaan, E., & Chun, E. (2018b). Syntactic Adaptation. In K. D. Federmeier & D. G. Watson (Eds.), *Psychology of Learning and Motivation* (Vol. 68, pp. 85–116). doi:10.1016/bs.plm.2018.08.003

Kaschak, M. P. (2007). Long-term structural priming affects subsequent patterns of language production. *Memory & Cognition, 35*(5), 925–937. doi:10.3758/bf03193466

Kaschak, M. P., & Borreggine, K. L. (2008). Is long-term structural priming affected by patterns of experience with individual verbs? *Journal of Memory and Language, 58*(3), 862–878. doi:10.1016/j.jml.2006.12.002

Kaschak, M. P., Kutta, T. J., & Jones, J. L. (2011). Structural priming as implicit learning: cumulative priming effects and individual differences. *Psychonomic Bulletin & Review, 18*(6), 1133–1139. doi:10.3758/s13423-011-0157-y

Kaschak, M. P., Loney, R. A., & Borreggine, K. L. (2006). Recent experience affects the strength of structural priming. *Cognition, 99*(3), B73-82. doi:10.1016/j.cognition.2005.07.002

Lawrence, M. A. (2016). *ez: Easy Analysis and Visualization of Factorial Experiments*. Retrieved from https://cran.r-project.org/package=ez

Ledoux, K., Traxler, M. J., & Swaab, T. Y. (2007). Syntactic priming in comprehension: evidence from event-related potentials. *Psychological Science, 18*(2), 135–143. doi:10.1111/j.1467-9280.2007.01863.x

Lenth, R. V., Buerkner, P., Herve, M., Love, J., Riebl, H., & Singmann, H. (2021). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. Retrieved from https://cran.r-project.org/package=emmeans

Leow, R. P. (2015). *Explicit Learning in the L2 Classroom: A Student-Centered Approach*. Retrieved from https://play.google.com/store/books/details?id=rKzABgAAQBAJ

Lum, J. A. G., Gelgic, C., & Conti-Ramsden, G. (2010). Procedural and declarative memory in children with and without specific language impairment. *International Journal of Language & Communication Disorders / Royal College of Speech & Language Therapists, 45*(1), 96–107. doi:10.3109/13682820902752285

McDonough, K., & Trofimovich, P. (2015). Structural priming and the acquisition of novel form-meaning mappings. In T. Cadierno & S. Wind Eskildsen (Eds.), *Usage-based perspectives on second language learning* (pp. 105–123).

Meara, P. M., & Rogers, V. E. (2019). *The LLAMA Tests v3. LLAMA B3 v3.00*. Cardiff: Lognostics.

Menenti, L., Gierhan, S. M. E., Segaert, K., & Hagoort, P. (2011). Shared language: overlap and segregation of the neuronal infrastructure for speaking and listening revealed by functional MRI. *Psychological Science, 22*(9), 1173–1182. doi:10.1177/0956797611418347

Monaghan, P., Ruiz, S., & Rebuschat, P. (2020). The role of feedback and instruction on the cross-situational learning of vocabulary and morphosyntax: Mixed effects models reveal local and global effects on acquisition. *Second Language Research*, 0267658320927741. doi:10.1177/0267658320927741

Montero-Melis, G., & Jaeger, F. T. (2020). Changing expectations mediate adaptation in L2 production. *Bilingualism: Language and Cognition, 23*(3), 602–617. doi:10.1017/S1366728919000506

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., … Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*(1), 195–203. doi:10.3758/s13428-018-01193-y

Peter, M. S., & Rowland, C. F. (2019). Aligning Developmental and Processing Accounts of Implicit and Statistical Learning. *Topics in Cognitive Science, 11*(3), 555–572. doi:10.1111/tops.12396

Pinker, S. (2009). *Language Learnability and Language Development*.

Plonsky, L., Marsden, E., Crowther, D., Gass, S. M., & Spinner, P. (2020). A methodological synthesis and meta-analysis of judgment tasks in second language research. *Second Language Research, 26*(4), 583–621. doi:10.1177/0267658319828413

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. Retrieved from R Foundation for Statistical Computing website: https://www.R-project.org/

Rebuschat, P. (2013). Measuring Implicit and Explicit Knowledge in Second Language Research: Measuring Implicit and Explicit Knowledge. *Language Learning, 63*(3), 595–626. doi:10.1111/lang.12010

Rebuschat, P., Monaghan, P., & Schoetensack, C. (2021). Learning vocabulary and grammar from cross-situational statistics. *Cognition, 206*, 104475. doi:10.1016/j.cognition.2020.104475

Segaert, K., Menenti, L., Weber, K., Petersson, K. M., & Hagoort, P. (2012). Shared syntax in language production and language comprehension--an FMRI study. *Cerebral Cortex , 22*(7), 1662–1670. doi:10.1093/cercor/bhr249

Shin, J.-A., & Christianson, K. (2012). Structural Priming and Second Language Learning. *Language Learning, 62*(3), 931–964. doi:10.1111/j.1467-9922.2011.00657.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

Stahl, A. E., & Feigenson, L. (2017). Expectancy violations promote learning in young children. *Cognition, 163*, 1–14. doi:10.1016/j.cognition.2017.02.008

Voeten, C. C. (2020). *buildmer: Stepwise Elimination and Term Reordering for Mixed-Effects Regression*. Retrieved from https://cran.r-project.org/package=buildmer

Walker, N., Monaghan, P., Schoetensack, C., & Rebuschat, P. (2020). Distinctions in the acquisition of vocabulary and grammar: An individual differences approach. *Language Learning, 70*(52). doi:10.1111/1467-923X.12837

Weber, K., Christiansen, M. H., Indefrey, P., & Hagoort, P. (2019). Primed From the Start: Syntactic Priming During the First Days of Language Learning. *Language Learning, 69*(1), 198–221. doi:10.1111/lang.12327

Wonnacott, E., Newport, E. L., & Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology, 56*(3), 165–209. doi:10.1016/j.cogpsych.2007.04.002

Yu, C., & Smith, L. (2007). Rapid Word Learning Under Uncertainty via Cross-Situational Statistics. *Psychological Science, 18*(5), 414–420. doi:10.1111/j.1467-9280.2007.01915.x

## Data, code and materials availability statement

All data, analysis code and materials used in this study can be found from the OSF repository for this study at https://doi.org/10.17605/OSF.IO/NKSU8 and from the IRIS database at https://www.iris-database.org/.

## Authorship and contributorship statement

Both authors of this paper meet the criteria for authorship set out by the International Committee of Medical Journal Editors and listed in Author Guidelines for *Language Development Research.* Specifically, the authors contributed to the following aspects: Giulia Bovolenta: experiment design, data collection, data analysis and interpretation, manuscript preparation. Emma Marsden: experiment design, data analysis and interpretation, manuscript preparation.

## Appendix S1 – Individual differences

We report here the results of the two cognitive tests that were administered: Serial Reaction Task and LLAMA B3. In order to determine whether the observed effects of our experimental manipulation were independent of any individual differences captured by these cognitive measures, we ran a series of *glmer* models where we added the z-transformed scores from cognitive measures as fixed predictors, in addition to the factors already entered in the main analysis. As random effects structure, we used the same structure that was originally used for the corresponding models in the main analysis. As in the main analysis, we used *buildmer* to simplify the models to only retain predictors that significantly improved model fit.

### Serial Reaction Task

There was no significant difference in mean SRT score between the two groups ($t(65.32) = -0.7601$, $p = 0.45$). The distribution of scores is shown in Figure 9. Both groups deviated from the normal distribution to some extent, which was significant in a Shapiro-Wilk test for the Control group ($W = 0.929$, $p = 0.03$) but not for the Surprisal group ($W = 0.953$, $p = 0.13$). When adding SRT scores to the model for accuracy on the Day 3 structure test block (new verbs), the effects of Group and Condition were still observed, in addition to a negative effect of SRT score (Table 4 & Figure 10). When adding SRT scores to the model for item endorsement in the GJT, SRT score was removed as a predictor during model selection as it had no significant effect on model fit, while the original Group x Verb Type x Grammaticality remained significant. We report the output of the initial model (with maximal fixed structure) for reference (Table 5).

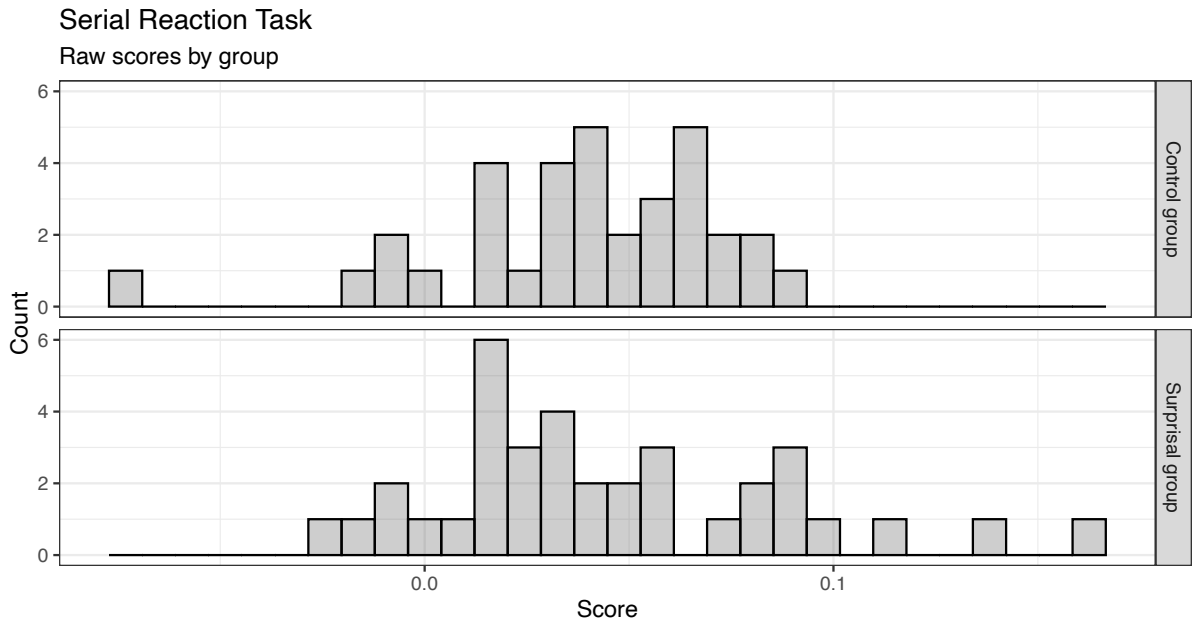


**Figure 9.** ***SRT scores by group***

**Table 4.** ***Final model for accuracy on Day 3 structure test (new verbs) with SRT score added to fixed effects structure***

| | Accuracy | | |
|---|---|---|---|
| *Predictors* | *Odds Ratios* | *CI* | *p* |
| (Intercept) | 1.14 | 0.67 – 1.94 | 0.622 |
| Structure (Active) | 3.09 | 1.57 – 6.05 | **0.001** |
| Group (Surprisal) | 2.65 | 1.42 – 4.95 | **0.002** |
| SRT score | 0.70 | 0.52 – 0.95 | **0.023** |
| Observations | 1120 | | |
| Marginal $R^2$ / Conditional $R^2$ | 0.102 / 0.483 | | |
| Random effects: (1 + Structure \| subject) | | | |

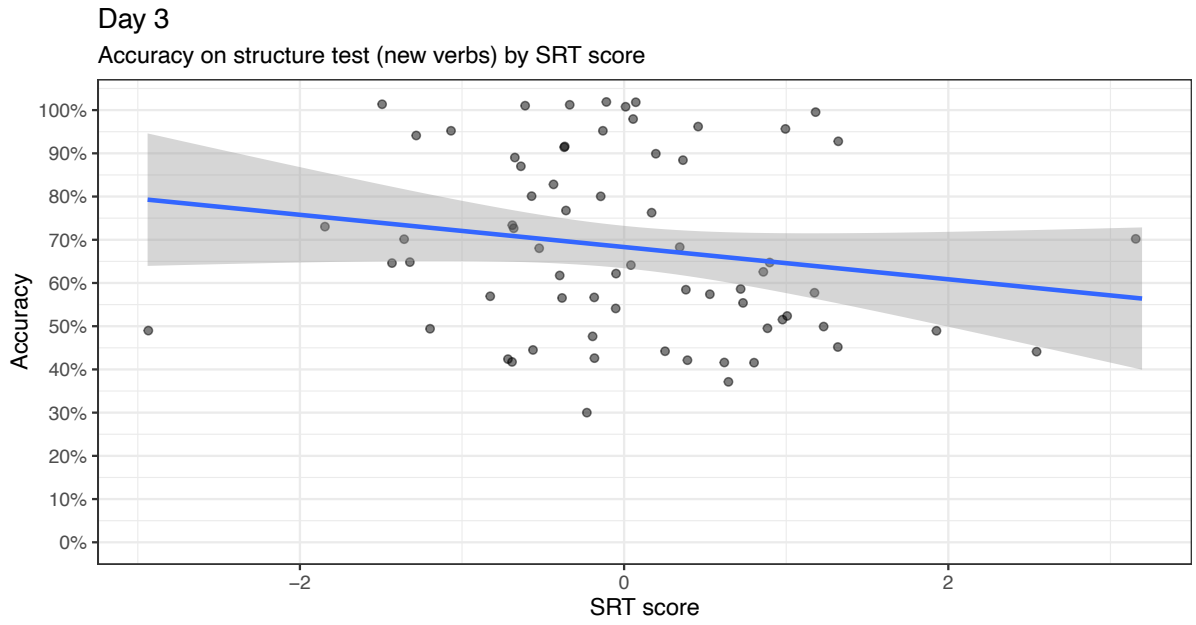


**Figure 10.** ***Accuracy on Day 3 structure test (new verbs) plotted against z-transformed SRT score***

**Table 5.** ***Initial model for endorsement in GJT, with SRT score added to fixed effects structure***

| | **Endorsement** | | |
|---|---|---|---|
| *Predictors* | *Odds Ratios* | *CI* | *p* |
| (Intercept) | 1.09 | 0.66 – 1.78 | 0.740 |
| cond [SG] | 0.51 | 0.25 – 1.02 | 0.056 |
| sentenceType [Grammatical] | 5.25 | 2.38 – 11.58 | **<0.001** |
| verbType [Active] | 4.60 | 2.23 – 9.50 | **<0.001** |
| SRT_z_score1 | 1.21 | 0.68 – 2.15 | 0.513 |
| cond [SG] * sentenceType [Grammatical] | 1.07 | 0.37 – 3.11 | 0.896 |
| cond [SG] * verbType [Active] | 0.75 | 0.28 – 2.01 | 0.571 |
| sentenceType [Grammatical] * verbType [Active] | 0.27 | 0.10 – 0.77 | **0.015** |
| cond [SG] * SRT_z_score1 | 0.89 | 0.44 – 1.82 | 0.750 |
| sentenceType [Grammatical] * SRT_z_score1 | 0.85 | 0.34 – 2.13 | 0.726 |
| verbType [Active] * SRT_z_score1 | 1.15 | 0.51 – 2.60 | 0.734 |
| (cond [SG] * sentenceType [Grammatical]) * verbType [Active] | 5.77 | 1.39 – 23.89 | **0.016** |
| (cond [SG] * sentenceType [Grammatical]) * SRT_z_score1 | 0.92 | 0.30 – 2.85 | 0.890 |
| (cond [SG] * verbType [Active]) * SRT_z_score1 | 0.96 | 0.35 – 2.65 | 0.939 |
| (sentenceType [Grammatical] * verbType [Active]) * SRT_z_score1 | 1.18 | 0.36 – 3.79 | 0.787 |
| (cond [SG] * sentenceType [Grammatical] * verbType [Active]) * SRT_z_score1 | 0.64 | 0.15 – 2.76 | 0.548 |
| Observations | 2240 | | |
| Marginal $R^2$ / Conditional $R^2$ | 0.179 / 0.464 | | |

Random effects: (1 + sentenceType + verbType + sentenceType:verbType | subject)

**LLAMA B3**

Due to a technical problem, we lacked LLAMA B3 scores for two participants. For remaining participants, there was no significant difference in mean LLAMA B3 scores between the two groups ($t(62.769) = -0.472$, $p = 0.64$). t = -0.47254). The distribution of scores is shown in Figure 11. Scores tended to deviate from normality, although this was only significant in the Control group ($W = 0.909$, $p = 0.01$) and not for the Surprisal group ($W = 0.970$, $p = 0.42$).

When adding z-transformed LLAMA B3 scores to the model for accuracy on the Day 3 structure test block (new verbs), all interactions were removed as they did not significantly improve model fit. The effects of Group and Condition were still observed, and the effect of LLAMA B3 score was not significant (Table 6). When adding LLAMA B3 scores to the model for item endorsement in the GJT, LLAMA B3 score was removed as a predictor during model selection as it had no significant effect on model fit, while the original Group x Verb Type x Grammaticality remained significant. We report the output of the initial model (with maximal fixed structure) for reference (Table 7).

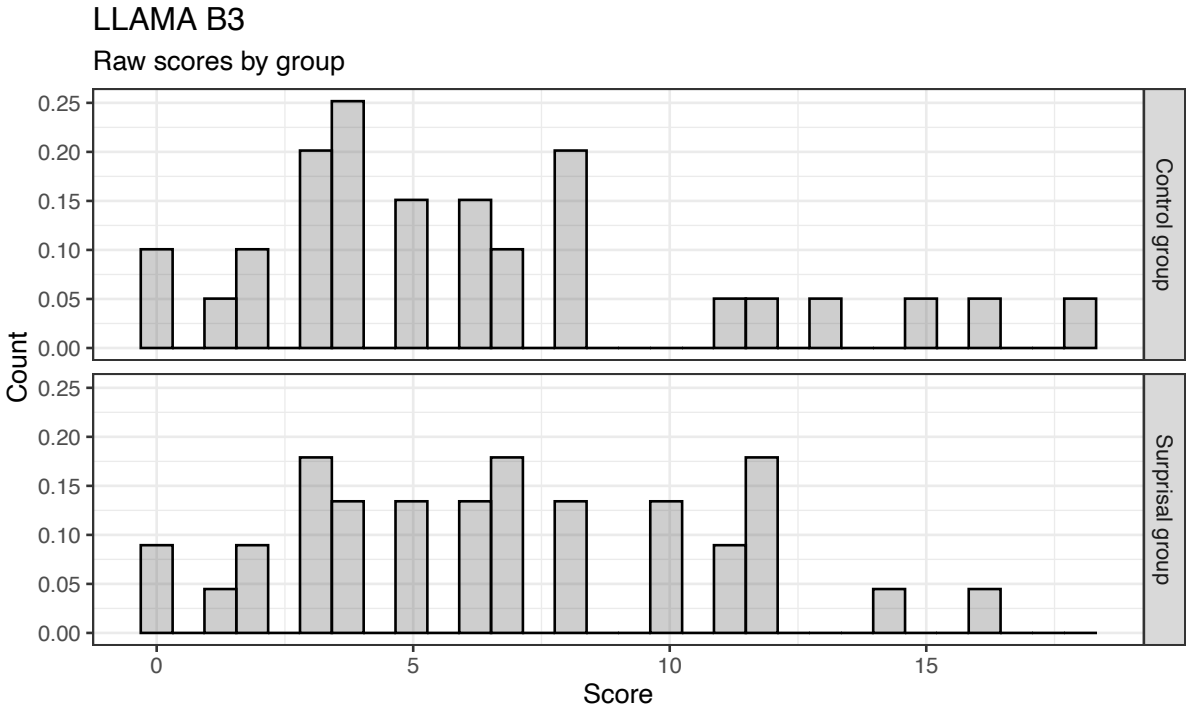


**Figure 11.** ***LLAMA B3 scores by group***

**Table 6.** ***Final model for accuracy on Day 3 structure test (new verbs) with LLAMA B3 score added to fixed effects structure***

| | Accuracy | | |
|---|---|---|---|
| *Predictors* | *Odds Ratios* | *CI* | *p* |
| (Intercept) | 1.08 | 0.64 – 1.83 | 0.779 |
| LLAMA B3 score | 1.10 | 0.79 – 1.53 | 0.569 |
| Structure (Active) | 3.48 | 1.84 – 6.58 | **<0.001** |
| Group (Surprisal) | 2.45 | 1.30 – 4.63 | **0.006** |
| Observations | 1088 | | |
| Marginal $R^2$ / Conditional $R^2$ | 0.098 / 0.464 | | |

Random effects: (1 + Structure | subject)

**Table 7.** ***Initial model for endorsement in GJT, with LLAMA B3 score added to fixed effects structure***

| | Endorsement | | |
|---|---|---|---|
| *Predictors* | *Odds Ratios* | *CI* | *p* |
| (Intercept) | 1.07 | 0.65 – 1.76 | 0.790 |
| cond [SG] | 0.52 | 0.26 – 1.05 | 0.067 |
| sentenceType [Grammatical] | 5.28 | 2.36 – 11.80 | **<0.001** |
| verbType [Active] | 4.11 | 1.99 – 8.47 | **<0.001** |
| LLAMA_B3_z_score | 0.99 | 0.62 – 1.60 | 0.979 |
| cond [SG] * sentenceType [Grammatical] | 1.02 | 0.35 – 2.98 | 0.971 |
| cond [SG] * verbType [Active] | 0.86 | 0.32 – 2.28 | 0.757 |
| sentenceType [Grammatical] * verbType [Active] | 0.30 | 0.11 – 0.88 | **0.028** |
| cond [SG] *LLAMA_B3_z_score | 0.57 | 0.28 – 1.15 | 0.116 |
| sentenceType [Grammatical] * LLAMA_B3_z_score | 0.77 | 0.37 – 1.61 | 0.489 |
| verbType [Active] * LLAMA_B3_z_score | 1.20 | 0.61 – 2.36 | 0.600 |
| (cond [SG] * sentenceType [Grammatical]) * verbType [Active] | 4.93 | 1.18 – 20.54 | **0.029** |
| (cond [SG] * sentenceType [Grammatical]) * LLAMA_B3_z_score | 2.64 | 0.90 – 7.70 | 0.076 |
| (cond [SG] * verbType [Active]) * LLAMA_B3_z_score | 1.17 | 0.44 – 3.14 | 0.753 |
| (sentenceType [Grammatical] * verbType [Active]) * LLAMA_B3_z_score | 0.76 | 0.29 – 1.97 | 0.574 |
| (cond [SG] * sentenceType [Grammatical] * verbType [Active]) * LLAMA_B3_z_score | 0.93 | 0.23 – 3.85 | 0.924 |
| Observations | 2176 | | |
| Marginal $R^2$ / Conditional $R^2$ | 0.185 / 0.466 | | |

Random effects: (1 + sentenceType + verbType + sentenceType:verbType | subject)

# Appendix S2 – Additional descriptive statistics

## Cross-situational learning task

Day 1, accuracy by block:

| Group | Block | Mean | Sd |
|---|---|---|---|
| Control | 1 | 0.51 | 0.13 |
| Control | 2 | 0.59 | 0.19 |
| Control | 3 | 0.62 | 0.18 |
| Control | 4 | 0.71 | 0.19 |
| Control | 5 | 0.71 | 0.19 |
| Control | 6 | 0.70 | 0.23 |
| Control | 7 | 0.65 | 0.17 |
| Surprisal | 1 | 0.53 | 0.11 |
| Surprisal | 2 | 0.59 | 019 |
| Surprisal | 3 | 0.63 | 0.22 |
| Surprisal | 4 | 0.69 | 0.22 |
| Surprisal | 5 | 0.71 | 0.20 |
| Surprisal | 6 | 0.73 | 0.20 |
| Surprisal | 7 | 0.68 | 0.21 |

Day 2, accuracy by block:

| Group | Block | Mean | Sd |
|---|---|---|---|
| Control | 1 | 0.66 | 0.18 |
| Control | 2 | 0.73 | 0.18 |
| Control | 3 | 0.80 | 0.17 |
| Control | 4 | 0.72 | 0.22 |
| Control | 5 | 0.73 | 0.15 |
| Control | 6 | 0.77 | 0.16 |
| Control | 7 | 0.77 | 0.17 |
| Control | 8 | 0.78 | 0.15 |
| Control | 9 | 0.82 | 0.16 |
| Control | 10 | 0.61 | 0.23 |
| Surprisal | 1 | 0.70 | 0.21 |
| Surprisal | 2 | 0.80 | 0.19 |
| Surprisal | 3 | 0.82 | 0.16 |
| Surprisal | 4 | 0.79 | 0.19 |
| Surprisal | 5 | 0.75 | 0.14 |
| Surprisal | 6 | 0.79 | 0.20 |
| Surprisal | 7 | 0.80 | 0.16 |
| Surprisal | 8 | 0.80 | 0.16 |

| | | | |
|---|---|---|---|
| Surprisal | 9 | 0.83 | 0.16 |
| Surprisal | 10 | 0.71 | 0.15 |

Day 3, accuracy by block:

| Group | Block | Mean | Sd |
|---|---|---|---|
| Control | 1 | 0.84 | 0.15 |
| Control | 2 | 0.66 | 0.20 |
| Control | 3 | 0.61 | 0.18 |
| Surprisal | 1 | 0.85 | 0.14 |
| Surprisal | 2 | 0.71 | 0.26 |
| Surprisal | 3 | 0.74 | 0.20 |

Day 3, accuracy on structure test block (New verbs), by structure:

| Group | Structure | Mean | Sd |
|---|---|---|---|
| Control | Passive | 0.52 | 0.26 |
| Control | Active | 0.71 | 0.28 |
| Surprisal | Passive | 0.65 | 0.34 |
| Surprisal | Active | 0.84 | 0.20 |

Vocabulary tests, accuracy by vocabulary item type:

| Group | Vocab type | Test | Mean | Sd |
|---|---|---|---|---|
| Control | Nountest | 1 | 0.65 | 0.27 |
| Control | Nountest | 2 | 0.70 | 0.23 |
| Control | Nountest | 3 | 0.74 | 0.25 |
| Control | Nountest | 4 | 0.85 | 0.23 |
| Control | Nountest | 5 | 0.90 | 0.17 |
| Control | Verbtest | 1 | 0.65 | 0.17 |
| Control | Verbtest | 2 | 0.62 | 0.27 |
| Control | Verbtest | 3 | 0.72 | 0.26 |
| Control | Verbtest | 4 | 0.80 | 0.18 |
| Control | Verbtest | 5 | 0.80 | 0.21 |
| Surprisal | Nountest | 1 | 0.70 | 0.23 |
| Surprisal | Nountest | 2 | 0.75 | 0.28 |
| Surprisal | Nountest | 3 | 0.81 | 0.25 |
| Surprisal | Nountest | 4 | 0.85 | 0.19 |
| Surprisal | Nountest | 5 | 0.86 | 0.19 |
| Surprisal | Verbtest | 1 | 0.67 | 0.26 |
| Surprisal | Verbtest | 2 | 0.65 | 0.23 |
| Surprisal | Verbtest | 3 | 0.77 | 0.20 |

| | | | | |
|---|---|---|---|---|
| Surprisal | Verbtest | 4 | 0.80 | 0.20 |
| Surprisal | Verbtest | 5 | 0.84 | 0.17 |

## Grammaticality Judgment Task

Grammaticality Judgment Task, endorsement by grammaticality and structure (verb type):

| Group | Sentence type | Error type | Verb type | Mean | Sd |
|---|---|---|---|---|---|
| Control | Ungrammatical | Actka | Active | 0.75 | 0.26 |
| Control | Ungrammatical | Passnoka | Passive | 0.51 | 0.28 |
| Control | Grammatical | None | Passive | 0.79 | 0.23 |
| Control | Grammatical | None | Active | 0.84 | 0.16 |
| Surprisal | Ungrammatical | Actka | Active | 0.61 | 0.35 |
| Surprisal | Ungrammatical | Passnoka | Passive | 0.39 | 0.29 |
| Surprisal | Grammatical | None | Passive | 0.69 | 0.26 |
| Surprisal | Grammatical | None | Active | 0.91 | 0.15 |

Grammaticality Judgment Task, *d'* scores (sensitivity to grammaticality) by structure (verb type):

| Group | Verb type | Mean | Sd |
|---|---|---|---|
| Control | Active | 0.44 | 1.24 |
| Control | Passive | 0.07 | 1.34 |
| Surprisal | Active | 0.33 | 1.37 |
| Surprisal | Passive | 0.01 | 1.54 |

## Appendix S3 – Additional figures (Day 2)

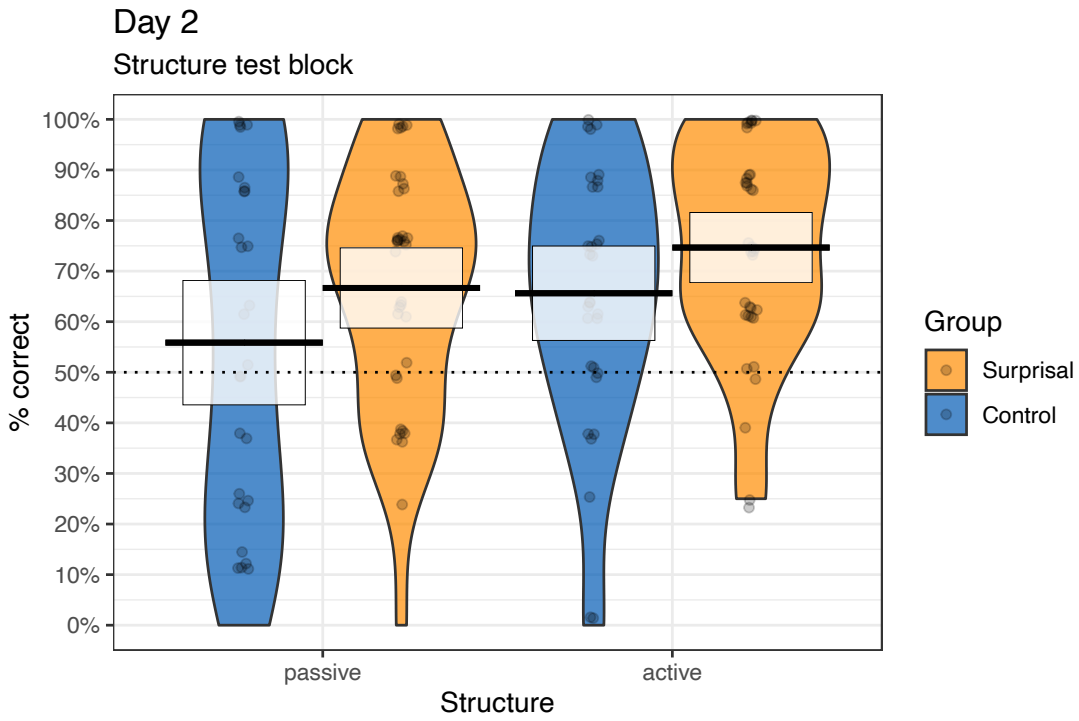


**Figure 12.** ***Accuracy on Day 2 structure test block, group means. Horizontal bars represent group means, shaded rectangles 95% CIs.***

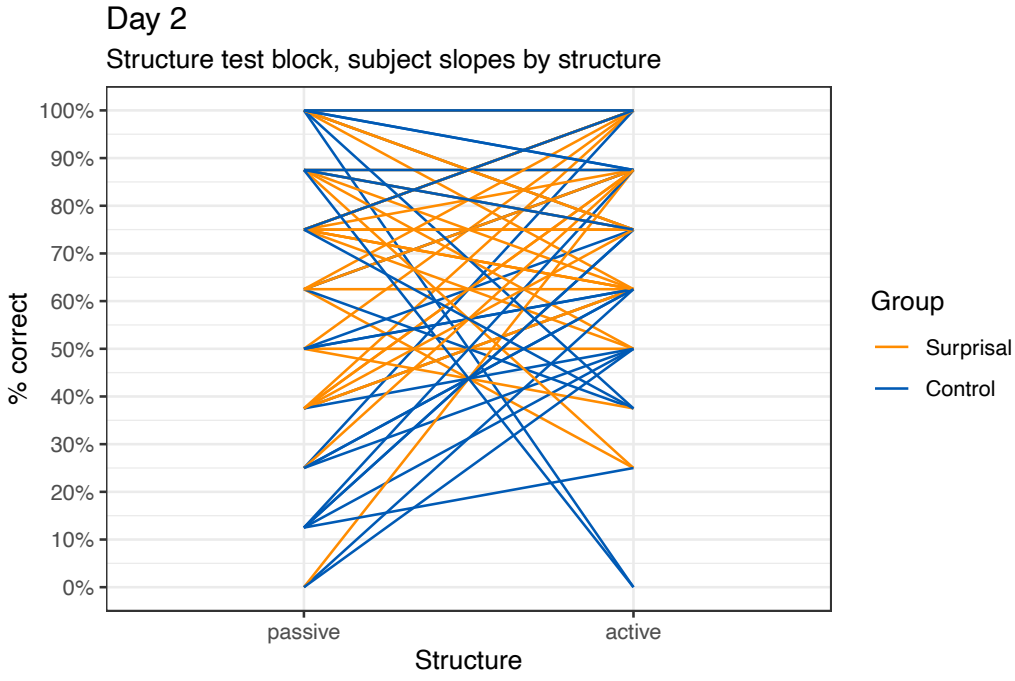


**Figure 13.** ***Accuracy on Day 2 structure test block by subject. Horizontal bars represent group means, shaded rectangles 95% CIs.***

## Appendix S4 – Details of Day 3 Structure Test block (Old verbs)

The Structure Test (Old verbs) block on Day 3 was affected by a counterbalancing problem, which meant roughly half of participants saw pictures described with the same structure they had encountered them with in the Day 2 Structure Test block, while the other half saw the pictures described with the other structure. This distinction was orthogonal to Group, although participants in the Surprisal group were numerically more likely to be exposed to the opposite structure, compared to the Control group (Table 1).

**Table 8.** ***Structure counterbalancing between Day 2 and Day 3 Structure Tests (Old verbs)***

| | Structure encountered on Day 3 | |
|---|---|---|
| | Same as Day 2 | Different |
| Control group | 19 | 15 |
| Surprisal | 15 | 21 |

Figure 1 shows the mean accuracy scores obtained by participants in the Structure Test (Old verbs) block on Day 3, broken down by whether the item had been seen in the Day 2 structure test block with the same structure. Figure 2 shows the same data, further broken down by whether participants had answered correctly (i.e., picked the correct structural interpretation) to the item on Day 2.

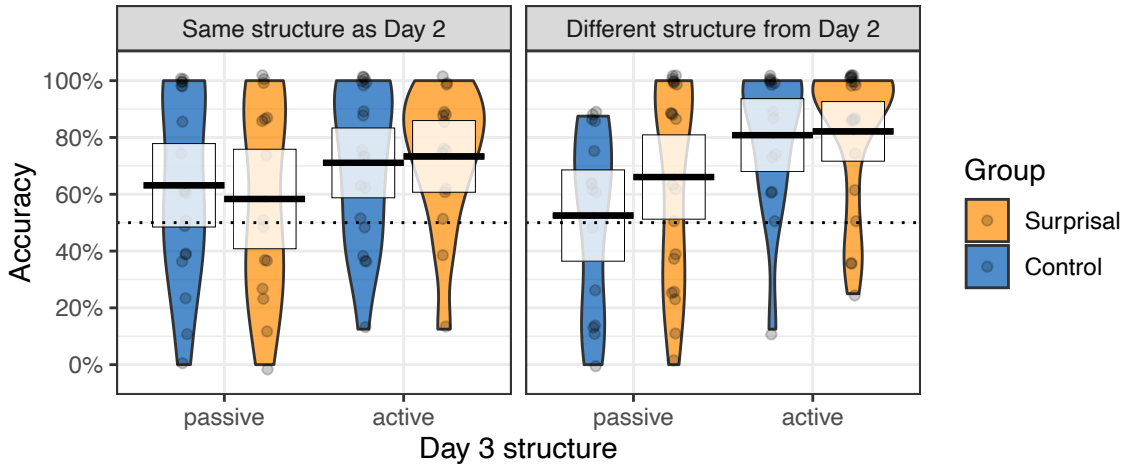


**Figure 14.** ***Mean accuracy on Day 3 Structure Test (Old verbs), divided by whether items used the same structure as in the Day 2 Structure Test.***

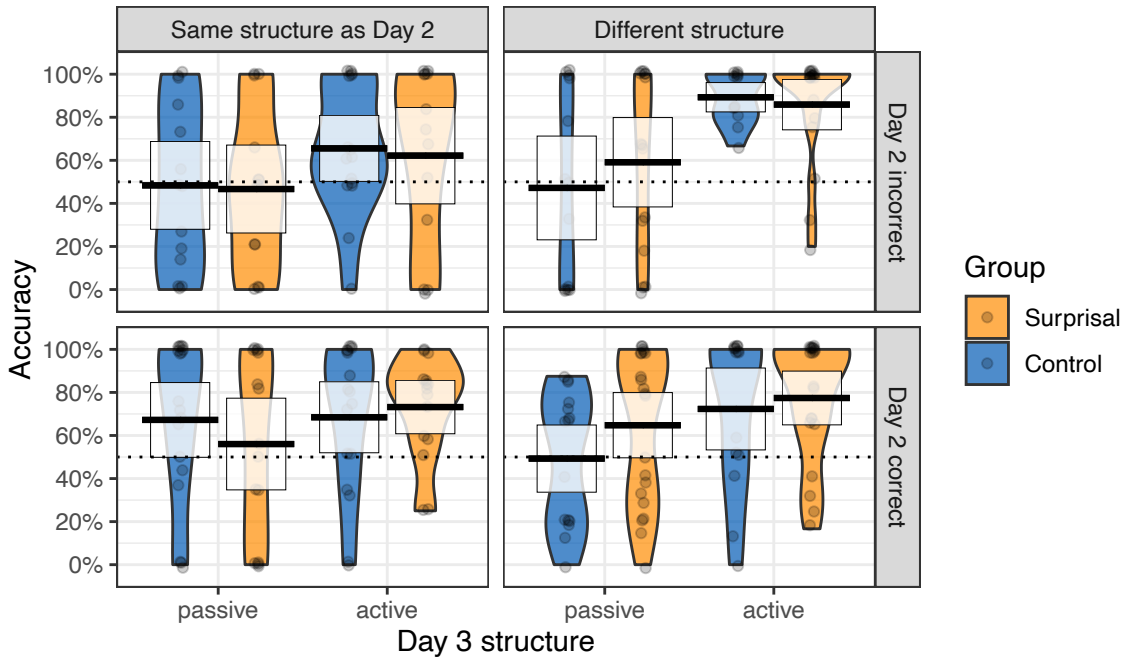


**Figure 15.** ***Mean accuracy on Day 3 Structure Test (Old verbs), divided by whether items used the same structure as in the Day 2 Structure Test, and by response accuracy to those items on Day 2.***

# Appendix S5 – Final statistical models for all tests

## Day 1, learning blocks

| | | Accuracy | | |
|---|---|---|---|---|
| *Predictors* | | *Odds Ratios* | *95% CI* | *p* |
| (Intercept) | | 2.36 | 1.90 – 2.94 | **<0.001** |
| Block | | 1.28 | 1.21 – 1.36 | **<0.001** |
| Observations | | 6720 | | |
| Marginal $R^2$ / Conditional $R^2$ | | 0.043 / 0.221 | | |
| Random effects: (1 + Block \| Subject) | | | | |
| *Groups* | | | *SD* | |
| Subject | (Intercept) | | 0.90 | |
| | Block | | 0.21 | |

## Day 2, learning blocks

| | | Accuracy | | |
|---|---|---|---|---|
| *Predictors* | | *Odds Ratios* | *95% CI* | *p* |
| (Intercept) | | 6.90 | 5.34 – 8.90 | **<0.001** |
| Block | | 1.12 | 1.06 – 1.18 | **<0.001** |
| Observations | | 5600 | | |
| Marginal $R^2$ / Conditional $R^2$ | | 0.013 / 0.256 | | |
| Random effects: (1 + Block \| Subject) | | | | |
| *Groups* | | | *SD* | |
| Subject | (Intercept) | | 1.01 | |
| | Block | | 0.13 | |

**Vocabulary test blocks: Verbs**

| | | **Accuracy** | |
|---|---|---|---|
| *Predictors* | *Odds Ratios* | *95% CI* | *p* |
| (Intercept) | 1.45 | 1.08 – 1.95 | **0.013** |
| Test | 1.33 | 1.24 – 1.41 | **<0.001** |
| Observations | 2800 | | |
| Marginal $R^2$ / Conditional $R^2$ | 0.037 / 0.230 | | |
| Random effects: (1 + Subject) | | | |
| *Group* | | *SD* | |
| Subject (Intercept) | | 0.91 | |

**Vocabulary test block: Nouns**

| | | **Accuracy** | |
|---|---|---|---|
| *Predictors* | *Odds Ratios* | *95% CI* | *p* |
| (Intercept) | 1.43 | 0.84 – 2.45 | 0.186 |
| Test | 1.69 | 1.51 – 1.88 | **<0.001** |
| Group (Surprisal) | 1.48 | 0.74 – 2.96 | 0.269 |
| Test x Group (Surprisal) | 0.80 | 0.69 – 0.93 | **0.004** |
| Observations | 2800 | | |
| Marginal $R^2$ / Conditional $R^2$ | 0.068 / 0.405 | | |
| Random effects: (1 + Subject) | | | |
| *Group* | | *SD* | |
| Subject (Intercept) | | 1.36 | |

**Day 2, Structure Test trials after feedback trials**

| | | Accuracy | |
|---|---|---|---|
| *Predictors* | *Odds Ratios* | *95% CI* | *p* |
| (Intercept) | 1.43 | 1.11 – 1.85 | **0.005** |
| Trial | 1.40 | 1.17 – 1.67 | **<0.001** |
| Observations | 1120 | | |
| Marginal $R^2$ / Conditional $R^2$ | 0.024 / 0.265 | | |
| Random effects: (1 + Trial \| Subject) | | | |
| *Groups* | | *SD* | |
| Subject | (Intercept) | 0.92 | |
| | Trial | 0.47 | |

**Day 2, Structure Test block**

| | | Accuracy | |
|---|---|---|---|
| *Predictors* | *Odds Ratios* | *95% CI* | *p* |
| (Intercept) | 2.03 | 1.34 – 3.07 | **0.001** |
| Structure (Active) | 1.44 | 0.86 – 2.39 | 0.162 |
| Observations | 1120 | | |
| Marginal $R^2$ / Conditional $R^2$ | 0.006 / 0.351 | | |
| Random effects: (1 + Structure \| Subject) | | | |
| *Groups* | | *SD* | |
| Subject | (Intercept) | 1.50 | |
| | Structure | 1.72 | |

**Day 3, Structure Test block (Old verbs)**

| | | Accuracy | |
|---|---|---|---|
| *Predictors* | *Odds Ratios* | *95% CI* | *p* |
| (Intercept) | 2.15 | 1.28 – 3.61 | **0.004** |
| Structure (Active) | 2.80 | 1.50 – 5.23 | **0.001** |
| Observations | 1120 | | |
| Marginal $R^2$ / Conditional $R^2$ | 0.040 / 0.506 | | |
| Random effects: (1 + Structure \| Subject) | | | |
| *Groups* | | *SD* | |
| Subject | (Intercept) | 1.89 | |
| | Structure | 1.87 | |

**Day 3, Structure Test block (New verbs)**

| | | Accuracy | |
|---|---|---|---|
| *Predictors* | *Odds Ratios* | *95% CI* | *p* |
| (Intercept) | 1.25 | 0.82 – 1.90 | 0.303 |
| Structure (Active) | 2.82 | 1.42 – 5.59 | **0.003** |
| Group (Surprisal) | 2.50 | 1.25 – 4.99 | **0.009** |
| Observations | 1120 | | |
| Marginal $R^2$ / Conditional $R^2$ | 0.073 / 0.496 | | |
| Random effects: (1 + Structure + Group \| Subject) | | | |
| *Groups* | | *SD* | |
| Subject | (Intercept) | 1.12 | |
| | Structure | 2.30 | |
| | Group | 1.59 | |

**Grammaticality Judgment Task: Endorsement**

| | Endorsement | | |
|---|---|---|---|
| *Predictors* | *Odds Ratios* | *95% CI* | *p* |
| (Intercept) | 1.05 | 0.66 – 1.67 | 0.845 |
| Sentence Type (Grammatical) | 4.45 | 2.19 – 9.03 | **<0.001** |
| Verb Type (Active) | 3.68 | 2.45 – 5.52 | **<0.001** |
| Group (Surprisal) | 0.52 | 0.27 – 1.00 | 0.050 |
| Sentence Type (Grammatical) x Group (Surprisal) | 1.15 | 0.43 – 3.05 | 0.781 |
| Verb Type (Active) x Group (Surprisal) | 0.85 | 0.49 – 1.48 | 0.562 |
| Sentence Type (Grammatical) x Verb Type (Active) | 0.40 | 0.22 – 0.73 | **0.003** |
| Sentence Type (Grammatical) x Verb Type (Active) x Group (Surprisal) | 4.44 | 1.85 – 10.64 | **0.001** |
| Observations | 2240 | | |
| Marginal $R^2$ / Conditional $R^2$ | 0.184 / 0.396 | | |
| Random effects: (1 + Sentence Type \| Subject) | | | |
| *Groups* | | *SD* | |
| Subject | (Intercept) | 1.14 | |
| | Sentence Type | 1.70 | |

**Grammaticality Judgment Task: *d'***

| Effects | *DFn* | *DFd* | *SSn* | *SSd* | *Generalised* $\eta^2$ | *F* | *p* |
|---|---|---|---|---|---|---|---|
| (Intercept) | 1 | 68 | 0 | 169.677 | 0 | 0 | 1 |
| Group | 1 | 68 | 4.514 | 169.677 | 0.017 | 1.809 | 0.183 |
| Verb Type | 1 | 68 | 0.301 | 89.125 | 0.001 | 0.230 | 0.633 |
| Group x Verb Type | 1 | 68 | 6.016 | 89.125 | 0.023 | 4.590 | **0.036** |

**Debriefing questionnaire**

| | Awareness | | |
|---|---|---|---|
| *Predictors* | *Odds Ratios* | *95% CI* | *p* |
| (Intercept) | 1.43 | 0.73 – 2.89 | 0.306 |
| Group (Surprisal) | 0.50 | 0.19 – 1.28 | 0.153 |
| Observations | 70 | | |
| $R^2$ Tjur | 0.029 | | |

## License