

Journal of Experimental Psychology: Learning, Memory, and Cognition

Access to Inner Language Enhances Memory for Events

Briony Banks and Louise Connell

Online First Publication, June 27, 2024. <https://dx.doi.org/10.1037/xlm0001351>

CITATION

Banks, B., & Connell, L. (2024). Access to inner language enhances memory for events.. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. <https://dx.doi.org/10.1037/xlm0001351>

Access to Inner Language Enhances Memory for Events

Briony Banks¹ and Louise Connell²

¹Department of Psychology, Fylde College, Lancaster University

²Department of Psychology, Maynooth University



Events are temporally bounded experiences involving people, objects, and actions that can be segmented into sequences of smaller, meaningful events (e.g., steps involved in constructing a piece of furniture), but the role of inner language in remembering such events has been unclear. We investigated whether inner language enhances memory for events in a naturalistic, nonverbal task where participants constructed simple models from memory. Across three experiments, we used linguistic suppression in a dual-task paradigm to test whether inner language improved overall memory performance and completion time, additionally exploring the number of events that could be recalled. We found that access to inner language at encoding consistently affected memory performance: when inner language was disrupted at encoding, participants were poorer at recalling the models and remembered fewer events. This effect was present whether or not the number of events to be recalled exceed event memory capacity (estimated as approximately seven to eight events). Critically, linguistic suppression impaired memory performance to a greater extent than a control secondary task that did not affect access to language; that is, impairment was not solely due to dual-task interference. The results support the proposal that inner language enhances event memory via a mechanism of linguistic bootstrapping, which makes event representation more efficient by allowing more information to be encoded in an event model even when language is not being used in the task. These findings therefore extend theories of event memory and add to a growing body of evidence that inner language is a highly valuable cognitive tool.

Keywords: event cognition, event memory, inner language, linguistic suppression, linguistic bootstrapping

We naturally carve up our conscious experience of life into events, such as driving to work, meeting friends for lunch or making a cake. These events are intrinsically temporal in nature—they are segments of time related to people, objects, and actions that are perceived to have a beginning and an end (Zacks, 2020; Zacks & Tversky, 2001). Empirical research has shown that events have internal structure (e.g., Zacks & Swallow, 2007), and that we can naturally segment coarse-grained events (e.g., making a cake) into connected, finer-grained events (e.g., weighing ingredients, mixing sugar and

butter, breaking eggs, etc.). Events therefore comprise rather complex and interconnected bundles of sensory and motor experiences that unfold sequentially over time to form a larger event.

Remembering complex sequences of events (e.g., the steps involved in making a cake) is a common part of everyday life, but due to their complexity recall is rarely perfect; that is, it likely surpasses memory capacity and cognitive resources to remember complex events fully and accurately. Accordingly, event memory has been proposed to rely on several mechanisms. Firstly, segmentation

Matthew Rhodes served as action editor.


Louise Connell  <https://orcid.org/0000-0002-5291-5267>


This work was supported by the European Research Council under the European Union's Horizon 2020 Research and Innovation Program (Grant Agreement 682848) to Louise Connell. The authors would like to thank Louise Brown for sharing materials for the visual patterns test, Nikoletta Alexandri and Emma Hewlett for their valuable help with video coding, and Regan Kelly for his valuable help with material creation and data collection for the visual patterns test. All materials, images, data, and code associated with this article are available at <https://osf.io/v8q47/> and licensed under a Creative Commons Attribution 4.0 International License (CC-BY), which permits the use, sharing, adaptation, distribution, and reproduction in any medium or format, as long as you give appropriate credit to the original authors and the source, provide a link to the Creative Commons license, and indicate if changes were made.


Open Access funding provided by Irish Research e-Library: This work is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0; <http://creativecommons.org/licenses/by/4.0>). This license

permits copying and redistributing the work in any medium or format, as well as adapting the material for any purpose, even commercially.

Briony Banks served as lead for data curation, formal analysis, investigation, and writing—original draft. Louise Connell served as lead for conceptualization, funding acquisition, methodology, project administration, supervision, and writing—review and editing and served in a supporting role for data curation and writing—original draft. Briony Banks and Louise Connell contributed equally to validation and visualization.

 The data are available at <https://osf.io/v8q47/>.

 The experimental materials are available at <https://osf.io/v8q47/>.

 The preregistered design is available at <https://aspredicted.org/hi9vw.pdf>, <https://aspredicted.org/y4av7.pdf>, <https://aspredicted.org/gc3yf.pdf>, <https://aspredicted.org/hf2gu.pdf>.

Correspondence concerning this article should be addressed to Louise Connell, Department of Psychology, Maynooth University, Maynooth, County Kildare, Ireland, or Briony Banks, Department of Psychology, Fylde College, Lancaster University, Bailrigg, Lancaster, LA1 4YF, United Kingdom. Email: louise.connell@mu.ie or brionbanks.psychology@gmail.com

of large events into smaller events is related to, and predicts, better event memory performance (Bailey et al., 2013; Flores et al., 2017; Sargent et al., 2013; Zacks et al., 2006; but see also Kurby & Zacks, 2011). Secondly, as an event is perceived, people form an online event model in working memory (Radvansky & Zacks, 2014), comprising a perceptual representation that can be maintained and updated as the event unfolds, particularly at event boundaries where prediction error spikes (see Zacks, 2020, for review). Thirdly, event memory can rely on existing long-term knowledge if a representation already exists; for example, if you have made a cake before, the long-term memory of this process can provide a scaffold for learning the steps involved in making a different cake (Radvansky & Zacks, 2014; Rubin & Umanath, 2015; Zacks, 2020).

Another potential mechanism for remembering events is through inner language. Also termed inner speech, verbal thinking and internal monologue (among other names), inner language can be defined as “the subjective experience of language in the absence of overt and audible articulation” (Alderson-Day & Fernyhough, 2015, p. 931). Vygotsky (1934/1986) first proposed that inner language develops in childhood alongside overt language and helps us to mediate our cognition and behavior. This initial theory of inner language has led to considerable focus on its potential role in executive functions, with studies often using linguistic suppression (i.e., overt verbal repetition of words or syllables) as a way to experimentally manipulate the contribution of language to a given task (see Nedergaard et al., 2023, for review). Linguistic suppression typically takes the form of continuously repeating an utterance such as “the” or “la” while performing a primary task of interest; if performance in the primary task suffers—and particularly if it suffers more than for an alternative secondary task—then it suggests that access to language is important for cognitive processing in the primary task. For example, continuously repeating “the” interferes with people’s ability to count the number of dots in an array more than does continuously tapping a finger, indicating that language plays a key role in counting processes (Logie & Baddeley, 1987). Fairly consistent evidence indeed demonstrates that inner language is used to support task switching (e.g., Baddeley et al., 2001; Emerson & Miyake, 2003; Saeki & Saito, 2004; for a review, see Cragg & Nation, 2010), with more tentative evidence that it supports inhibitory control in nonverbal reasoning (e.g., Dunbar & Sussman, 1995; Wallace et al., 2017). However, the role of language is theorized to extend broadly across cognition (Borghini et al., 2019; Connell, 2019; Dove, 2020; Louwerse, 2011; Wingfield & Connell, 2022), and accordingly, a role for inner language has been demonstrated in domains such as categorization (He et al., 2019; Lupyan, 2009; Roberson & Davidoff, 2000; Winawer et al., 2007), learning novel categories (e.g., Minda et al., 2008; Zeithamova & Maddox, 2007), abstract word processing (Fini et al., 2022), metacognition (for a review, see Morin, 2018), and mental arithmetic (e.g., Frank et al., 2012; Imbo & LeFevre, 2010; Logie et al., 1994; Robert & LeFevre, 2013; Seitz & Schumann-Hengsteler, 2002; Trbovich & LeFevre, 2003).

Most pertinent to the present study is the role of inner language in memory. Inner language is well established to be a part of working memory in the form of the phonological loop for storage and rehearsal of verbal information (Baddeley & Hitch, 1974), much evidence for which comes from linguistic and verbal tasks (e.g., the phonological similarity effect, Baddeley, 1966). However, inner language can also facilitate nonverbal memory, for example, memory for previously learnt visual images (Brandimonte et al., 1992a, 1992b; Pelizzon et al., 1999),

faces (Nakabayashi & Burton, 2008), and even reproduction (i.e., free recall) of a complex line drawing (Bek et al., 2009). Most recently, Dymarska et al. (2022) found that participants’ ability to remember sequences of pictured objects (e.g., ingredients for a recipe) was impaired when they performed linguistic suppression while encoding the objects, even though the task itself was nonverbal. That is, there is reasonably consistent evidence that inner language facilitates memory for a range of discrete input types (e.g., verbal words, nonverbal faces, and objects).

When it comes to event memory, however, the role of inner language is less clear. Much research on the role of language in event memory has come from the perspective of linguistic relativity. Because different languages grammatically mark event structures in different ways, such as including information about path, manner, or aspect in motion verbs, many studies have investigated concomitant cross-linguistic differences in event representations (Athanasopoulos & Bylund, 2013; Flecken et al., 2015; Santin et al., 2021). Nonetheless, evidence for such linguistic relativity effects is inconsistent, particularly in nonverbal tasks that do not involve overt language (e.g., Papafragou et al., 2002; Skordos et al., 2020; ter Bekke et al., 2022). However, the potential ability of inner language to facilitate event memory goes beyond linguistic relativity effects. Rather than investigating differences between languages, an alternative perspective comes from investigating whether the presence of language itself—regardless of what that language might be—affects cognition and memory. A limited number of studies using linguistic suppression suggest, albeit inconclusively, that availability of inner language may facilitate remembering action events. Jaroslawska et al. (2018) examined memory for sequences of verbal instructions for simple manual actions using everyday objects (e.g., “pick up the blue pencil”) and found that linguistic suppression impaired accuracy of action reproduction in two out of three experiments, but to a lesser extent than a motor interference task involving repetitive hand actions. As such, rather than demonstrating a clear role for language in event representations, the effect of linguistic suppression could be explained by the extra demands of a secondary task (i.e., dual-task interference). Studies using demonstrated actions rather than verbal instructions have found mixed effects. Mitsuhashi et al. (2018) found that participants were less accurate at reproducing named hand gestures when linguistic suppression was performed during encoding, and that the effects were greater than when a concurrent motor/spatial task (finger tapping) was performed. However, in a similar task, Gimenes et al. (2016) observed that although linguistic suppression led to poorer gesture reproduction, performance was only equivalent to a motor interference condition, suggesting the results could be explained by an overall dual-task effect (see also Trueswell & Papafragou, 2010, for comparable results regarding memory for motion events). Thus, the event memory literature is currently conflicted as to whether language plays a role in event memory, with the available evidence suggesting that any robust linguistic effects are most likely restricted to circumstances where language is overtly being used during the task. It therefore remains to be seen whether inner language may actually benefit memory for nonverbal events as it does for simpler nonverbal stimuli such as objects and faces.

The Linguistic Bootstrapping Hypothesis

We propose that inner language can enhance memory for events specifically through a mechanism called linguistic bootstrapping

(Connell & Lynott, 2014). We routinely attach linguistic labels (i.e., words and phrases) to the rich bundles of sensorimotor information that make up concepts; for example, the sensorimotor referent of the word “cake” is a soft, sweet, edible object. When there are insufficient cognitive resources to maintain a full sensorimotor representation in memory, such as when many concepts are involved or the representation itself is highly complex or detailed, then a linguistic label can replace a portion of the sensorimotor representation to free up resources. In this way, a word or phrase can act as a cognitively efficient placeholder in memory for its sensorimotor referent because linguistic labels take up less representational “space” and cognitive resources than their sensorimotor counterparts (Dymarska et al., 2022; see also Barsalou et al., 2008; Connell, 2019; Louwerse, 2011). By preserving the structure of the representation while freeing up resources to extend or manipulate it as required, a linguistic label can therefore bootstrap complex mental representations that would otherwise outstrip available capacity. Encoding complex sensorimotor information, such as that involved in events, may substantially benefit from this linguistic bootstrapping mechanism. For example, making a cake might involve remembering a complex sequential combination of visual, tactile, olfactory, and gustatory information as well as hand and arm movements (i.e., the core ingredients, tools, and actions needed to perform the event), as well as more complex multisensory information such as the changing consistency of the cake mixture at different stages, spatial information and quantity. Using linguistic labels in place of some of this complex sensorimotor information—for example, “two eggs,” “beat,” “stiff”—may therefore use fewer cognitive resources than representing their rich sensorimotor referents, and thus enhance the quality and quantity of event representations that can be recalled from memory. Specifically, we propose that inner language is routinely used to help us remember and reproduce events even without linguistic input (i.e., when the events are completely nonverbal).

Studies demonstrating a role for inner language in memory for a variety of nonverbal tasks (as described earlier) provide preliminary evidence that the linguistic bootstrapping mechanism may be used to support event memory. However, none specifically tested the linguistic bootstrapping hypothesis or memory for events, and mixed results and methodological differences mean that the specific role of language in facilitating memory for sequences of actions (a core part of event memory) is still unknown. Moreover, most of the memory studies discussed so far have only tested the role of inner language during encoding. The linguistic bootstrapping hypothesis predicts that when remembering events, inner language should be beneficial at both encoding and recall: it is used during encoding to create efficient placeholders for sensorimotor representations, but can also be relevant during recall when linguistic labels may serve as cues to “flesh out” full sensorimotor representations. It is therefore important to test the contribution of inner language at both stages of the memory process to determine its full role in event memory.

The Present Study

Across three preregistered experiments, we investigated whether inner language plays a critical role in memory for events. Specifically, we tested the linguistic bootstrapping hypothesis that having inner language available would enhance event memory by allowing more efficient representation of event information, even when

language is not involved in or relevant to the task. We used a naturalistic, nonverbal event reproduction task (reconstructing physical models from memory) to represent “large” events encountered in everyday life that can be segmented into smaller events (e.g., remembering the steps needed to construct a piece of furniture). We tested use of inner language at both encoding (learning) and recall (event reproduction) by using linguistic suppression as a secondary task. In line with linguistic bootstrapping, we expected that access to inner language would enhance overall task performance: specifically, participants would be able to reconstruct the models from memory more accurately and more quickly when they could use linguistic labels as placeholders in complex event representations. In terms of our experimental manipulations, we hypothesized that when linguistic suppression was performed at either encoding or recall, and use of inner language was impaired, participants would exhibit poorer construction of the model from event memory. The study comprised four experiments: in Experiment 1, participants constructed a model from memory with linguistic suppression manipulated between groups. Experiment 2 refined the design of Experiment 1 and additionally tested whether effects of inner language were stronger for longer sequences of events where memory capacity was particularly strained: Experiment 2a first estimated memory capacity for events by comparing different sequence lengths, and Experiment 2b then tested the role of inner language for sequence lengths that were within and beyond memory capacity. Experiment 3 then tested whether the effects observed in Experiments 2a and 2b were truly due to use of inner language or whether they were dual-task interference effects, by using a secondary control task in comparison with linguistic suppression.

The study is novel in several respects. To the best of our knowledge, it is the first to study memory for naturalistic events via nonverbal event reproduction, using a large event that can be segmented into smaller events. We developed a novel event memory task (reconstructing models after viewing a video of their construction) that would test memory for complex sensorimotor events where participants have to remember (and reproduce) multisensory, spatial, and action information in a sequential manner. The task enables free recall to be tested rather than recognition memory, and allows use of inner language to be manipulated at both encoding and recall. In Experiment 3, we developed a novel control task as a comparison for linguistic suppression, suitable for use in tasks requiring motor processes. Lastly, we used the results of the studies to estimate average memory capacity for events which, to the best of our knowledge, was previously unknown.

Experiment 1: Does Language Enhance Memory for Events?

In this experiment, we tested whether inner language contributes to a naturalistic event memory task: constructing a simple wooden birdhouse. We asked participants to watch an instructional video that showed the incremental construction of a simple birdhouse model, and then immediately asked them to reconstruct the model from memory using the same components. Critically, we manipulated participants’ use of inner language during the task by asking them to perform linguistic suppression at encoding (i.e., when watching the instructional video) and/or recall (i.e., when constructing the model). We recorded participants during model construction and measured their overall memory performance (i.e., their completeness and correctness in reproducing each step of the model’s

construction), and their latency of performance (i.e., how long it took them to reproduce the model).

We predicted that memory performance would be poorer when use of inner language was limited due to linguistic suppression, because participants' initial encoding of event representations would not properly benefit from linguistic bootstrapping (i.e., when they performed linguistic suppression at encoding) and because participants could not properly access linguistic labels in their memory of events (i.e., when they performed linguistic suppression at recall). We also predicted that performance would be worst when inner language had been available at encoding and was then suddenly unavailable at recall (i.e., an interaction in the timing of linguistic suppression). That is, losing access to encoded linguistic representations at the point of recall could disrupt event memory more than simply encoding the events without language in the first place.

Method

The experiment's design and hypotheses were preregistered at <https://aspredicted.org/hi9vw.pdf>; all methods and analyses follow the preregistration unless otherwise specified. Materials, data, code, and full results output are available at <https://osf.io/v8q47/> (Connell & Banks, 2024).

Participants

Ninety-two participants took part in the study, recruited from Lancaster University for payment or course credit. Twelve participants were excluded for not meeting preregistered inclusion criteria: six failed to fully complete at least one step of the model, and six participants were nonnative speakers of English.¹ The final sample size for analysis was 80 participants who were all native speakers of English (65 female; $M_{\text{age}} = 20.52$ years, $SD = 4.79$). Sample size was determined via sequential hypothesis testing with Bayes factors (BF; Schönbrodt et al., 2017), where N_{min} was set at 40 (10 participants per condition), and recruitment continued until the interaction between encoding and retrieval cleared the prespecified grade of evidence $BF \geq 5$ (or its reciprocal 1/5) for both performance score and completion time (i.e., there was evidence for or against the Step 3 over the Step 2 model—see Statistical Analysis section) or until we reached N_{max} of 80 (20 participants per condition).² Evidence against the interaction was present at $N = 76$ for completion time but not for performance score, and so recruitment continued until N_{max} .

Ethics and Consent

Two consent forms were used for this study. The first contained standard terms of informed consent relating to participation in the experiment, data collection, and the sharing of all anonymized, alphanumeric data in a public data repository. The second form specifically related to release of audiovisual recordings and asked participants whether they consented to public sharing of their videos (which included archiving in a public data repository). Three options were available: to decline to share their video publicly, to consent to share their video on condition that their face was masked (blurred), or to consent to share their video with their face visible. In total, 76 out of 80 participants consented to share their video recordings (43 masked and 33 with face visible) and only four declined. Blank copies of the consent forms are available at <https://osf.io/v8q47/>, and all data have been shared on the Open Science Framework (OSF) in accordance

with participants' individual consent choice (i.e., video data are only available for the 76 participants who opted to publicly share their videos). This and subsequent experiments reported in the present study received ethical approval from the Lancaster University Faculty of Science and Technology Research Ethics Committee (application code FST17003).

Materials

For the event memory task, we used a simple wooden birdhouse craft kit (see Figure 1) comprising 20 individual parts and measuring approximately 18×13 cm. The kit was marketed as being suitable for construction by children aged six and upward, and could be constructed by hand by slotting the pieces together without adhesives or tools. We recorded a video of a demonstrator constructing the birdhouse, to be used as both an instructional video in the main experiment and to determine the event boundaries for later video coding (see Segmentation Task section). The instructional video started with the individual pieces of the birdhouse laid out clearly on a table and the demonstrator sitting behind the table with only upper body and arms visible (the head and face were obscured) so that the main focus of attention would be the pieces of the birdhouse and their manual construction. The video included sounds of the pieces fitting together but there was no dialogue or background noise, and was 3 min 10 s long. The video was recorded using an iPad (sixth generation) rear camera at 1080p and 30fps, edited using iMovie (Version 10.2.5), and exported as a QuickTime movie in MP4 format at 1080p (1920 \times 1080 pixels) resolution.

Segmentation Task

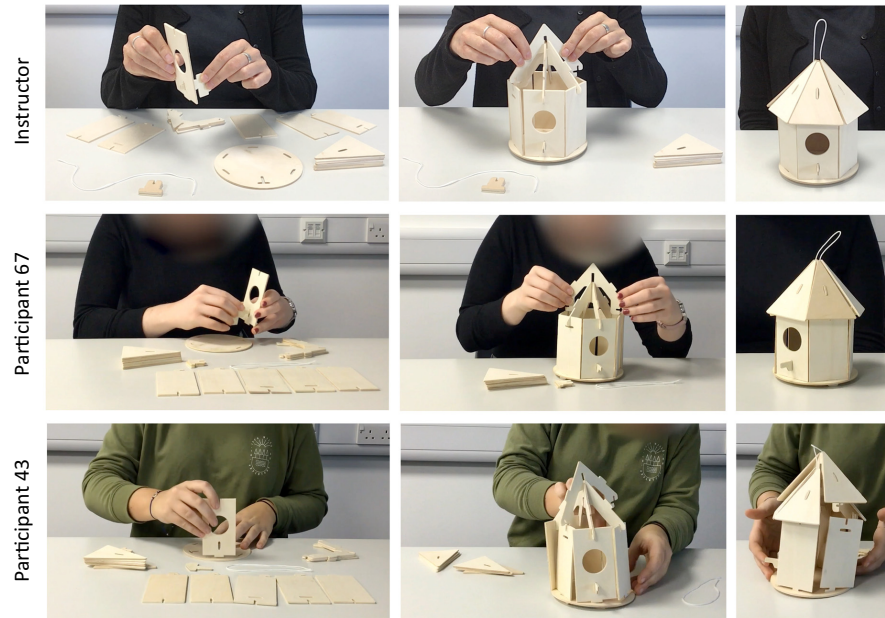
In order to score reconstruction performance separately for each "step" of building the model, we first had to determine what those steps were. Rather than use the steps provided in the original model instructions, or base the steps on experimenter intuition, we opted to establish the steps empirically based on the natural event boundaries present in the birdhouse construction process. We therefore carried out an event segmentation task (e.g., Zacks & Swallow, 2007) with a small group of participants who did not take part in the main experiments ($N = 9$, eight female; $M_{\text{age}} = 21.00$ years, $SD = 3.46$). Methods and results are summarized here but are fully reported as additional online materials (<https://osf.io/v8q47/>). We used two shorter practice videos of different activities (also available as additional online materials) to ensure that participants understood the task and could carry it out accurately prior to segmenting the instructional video for the birdhouse. All videos showed the same framed shot of the demonstrator's upper body and desk displaying the activity. The procedure closely followed that of previous event

¹ These excluded participants took part for psychology course credit which (due to university policy) meant they could not be prescreened on the basis of native language; they were therefore allowed to complete the study despite not meeting the eligibility criteria but their data were discarded. This criterion and procedure was followed in each experiment.

² Our N_{min} and N_{max} sample sizes were determined based on experimental design (i.e., allowing between 10 and 20 participants per cell of the design), logistical feasibility (i.e., time required for testing and coding the videos of up to 80 participants), and what we considered to be a sufficient test of the hypothesis (i.e., an effect that was still equivocal after testing 80 participants might be too small to offer convincing support for the theory).

Figure 1

Early, Middle, and Final Stages of Construction of the Birdhouse Model in Experiment 1 Taken From the Instructional Video (Top Row), a High-Performing Participant (Middle Row), and a Poor-Performing Participant (Bottom Row)



Note. The top-right image shows the correctly constructed model. Full videos can be viewed on OSF; participant numbers relate to Experiment 1 only. OSF = Open Science Framework. See the online article for the color version of this figure.

segmentation tasks (e.g., Newton, 1976; Zacks et al., 2001). Participants were instructed to identify the different meaningful units that they thought made up the activity shown in each video by pressing the spacebar whenever an event ended, and saying out loud what they saw happening. Participants wore headphones and a lapel microphone throughout the task, and key press and verbal responses were recorded for analysis. Videos were always presented in the same order (Practice 1, Practice 2, birdhouse).

Participant data were analyzed to identify the most commonly perceived event boundaries across participants. We matched the timing of event boundaries (i.e., each key press) identified by each participant to their corresponding verbal responses (e.g., Participant 1 pressed space bar at 18 s, which corresponded with the verbal response “She attaches this component to a circular piece of wood”). Based on these responses, we identified all meaningful events perceived by participants and labeled them according to the corresponding actions occurring in the video. Less meaningful segments (e.g., the demonstrator picking up pieces) were not included in the analysis. We considered an event boundary to be valid for our purposes if it was identified by at least seven out of nine participants. Completion of six critical events met this criteria with a very high level of overall agreement (90.7%): (a) the door and base; (b) walls; (c) roof beams; (d) structural supports; (e) string; and (f) roof tiles. We therefore considered these six events to represent reasonable approximations of the steps involved in model reconstruction, and used them as the basis of coding performance in Experiment 1.

Design and Procedure

Availability of inner language was manipulated between-participant using linguistic suppression as a secondary task. Linguistic suppression has been used widely in cognitive psychology to study inner language in a range of tasks and behaviors (see Nedergaard et al., 2023, for a review), particularly in relation to memory (e.g., Baddeley, 1966; Hitch et al., 1995; Jaroslawska et al., 2018), and is known to specifically disrupt linguistic processing (i.e., inner language) over and above general executive or attentional processes (Larsen & Baddeley, 2003). The task involves repeating words or syllables, thus involving linguistic and articulatory processes in a relatively simple and executive undemanding task. We chose to use the word “the” in the task to ensure that participants were repeating a real word (i.e., with practiced articulation), but one that is empty of meaning without context. Participants were randomly allocated to one of four conditions: no linguistic suppression, linguistic suppression at encoding (i.e., performed when watching the instructional video), linguistic suppression at recall (i.e., performed when constructing the model), or linguistic suppression at both encoding and recall.

Participants were first asked to read the study information sheet and to sign both consent forms. They then read the following instructions displayed onscreen:

You are going to watch a video of someone putting together a wooden model. When the video has finished you will be asked to put the model together yourself, in exactly the same way and in exactly the

same order. There is no time limit for putting it together but we'd like you to do it as accurately as possible, exactly as it is shown in the video. Please pay close attention to the video as once it has finished you will not be able to see it again!

For participants carrying out linguistic suppression during encoding, the experimenter (Briony Banks) then explained that they would be required to say the word “the” out loud repeatedly while watching the video. Participants were asked to say “the” at a rate of approximately two words per second (e.g., Larsen & Baddeley, 2003), which they practiced along with the experimenter and on their own for a few seconds, until the experimenter was happy that they were performing it correctly. The experimenter told them that if at any time they slowed down or stopped saying “the,” that she would prompt the participant to start again. All participants were asked to put on the headphones (Steelseries 5H V2 USB gaming headset) to watch the video. Participants carrying out linguistic suppression were asked to start saying “the” before the experimenter played the video, and told to stop as soon as the video ended. The video and instructions were played via PsychoPy2 (Version 1.90.2; Peirce et al., 2019) played at full resolution on a 24-in. 1920 × 1080 desktop monitor, which was the same setup as in the segmentation task. While the video was playing, the experimenter observed the participant to ensure that they watched the video and did not do anything additional that might interfere with the task (e.g., tapping their hands or feet).

After the video had finished playing, the participant sat at a separate table to build the model from memory. The pieces of the birdhouse were laid out on a table in exactly the same configuration for each participant (see Figure 1), but were initially hidden from view by a box covering all pieces. Participants were reminded that they needed to construct the model in exactly the same way that they had seen in the video and that there was no time limit, and they were asked to let the experimenter know when they had finished the model or could not get any further. Participants who were performing linguistic suppression were asked to repeat “the” for the whole time that they were building the birdhouse, and to stop once they had finished. If they had not performed linguistic suppression while watching the video, they were then instructed how to do so and asked to practice for a few seconds to make sure they could do it correctly. To control for this extra instruction and verbalization before the memory task in only one of the participant groups, in all other groups the experimenter pretended to check the camera setup and asked the participant to recite the months of the year while she did this. For all participants, the experimenter then started the video recording, and participants performing linguistic suppression were asked to start saying “the.” The experimenter removed the box covering the birdhouse pieces and told the participant that they could begin. Videos were recorded using an iPad (sixth generation) mounted on a tripod, using the rear camera at 1080p and 30fps, framing the desk and participant’s torso and head so that the birdhouse and participant’s upper body were visible.

While each participant constructed the birdhouse, the experimenter sat at a desk visible to the participant but with her back to them, so that she could hear the participant without distracting them from the task. In all linguistic suppression conditions, if the experimenter noticed the participant slowing for several utterances (i.e., a clear reduction in the rehearsed rate of approximately two “the” per second that persisted for 3–4 s) or if the participant stopped

saying “the” completely for the duration of several utterances (i.e., missed 3–4 “the” in a row), she intervened to correct them. Interventions did not happen often but, where they were necessary, they took the form of the experimenter clearly and loudly repeating “the” at the rehearsed rate of two per second, which prompted the participant to continue correctly (i.e., echoing the original training in linguistic suppression). When the participant declared they had finished, the experimenter then stopped the recording and asked the participant several debrief questions (how had they found the task? Had they used any words to help them construct the birdhouse or had they just relied on what they had seen? How had the linguistic suppression affected them?). They were then asked to provide basic demographic information to the experimenter, given a debrief sheet and compensated accordingly.

Data Preparation and Analysis

Video Coding and Scoring. Participant performance was measured by scoring their construction of each model step (i.e., each of the six smaller events identified in the segmentation task) in the video recordings. We first developed a coding scheme (available at <https://osf.io/v8q47/>) to score memory performance in each step based on four preregistered criteria: completion, errors, multiple attempts, and serial order. Completion scored whether a given step (e.g., the birdhouse roof) was present and complete without errors or pieces missing in the final version of the model; only the final state of the model was considered in this criterion, and any failed attempts or errors made earlier in construction were not taken into account. Errors scored whether any errors or deviations from the instructional video were made at any time during construction of a given step, even if they were later corrected. Multiple attempts scored whether there were any failed attempts at constructing the step (e.g., dismantling fitted pieces to try an alternative or repeatedly trying to fit incorrect pieces in the same position); this criterion was included to penalize participants for using trial and error to perform the task). Participants could score either a full point (1), half a point (0.5), or 0 for each of these three criteria, with higher scores reflecting better performance; however, if a step was never attempted (i.e., skipped entirely in model construction), participants scored –1 on the multiple attempts criterion. Finally, serial order scored whether the steps were constructed in the correct order or sequence as per the instructional video, based on all attempts of each step during the whole task, and was calculated as Levenshtein distance (computed using the stringdist package, van der Loo, 2014, in R Studio), reversed with a floor of zero so higher scores reflected better performance. Video recordings were coded and scored using ELAN software (ELAN, 2019). Each event was identified, labeled, and scored in ELAN following the specific definitions and criteria in the coding scheme.

Videos were first coded and scored by the main experimenter. To ensure objectivity, a second coder who was blind to the study aims and hypotheses subsequently coded and scored a sample of 17 videos (22% of all participants) using the same coding scheme and procedure. The sample was pseudo-randomly selected to include an approximately equal number from each of the four conditions, and to cover a systematic cross-section of summed scores. Following training on two videos not included in the sample, the second coder independently scored the videos; general questions about the coding scheme were allowed but not specific questions about participant behavior. Intercoder agreement on each of the four

scoring criteria (errors, completion, multiple attempts, and serial order) was analyzed using weighted Cohen’s kappa (GraphPad Software, n.d.) due to the graded nature of the scores (e.g., 0 and 0.5 are closer in agreement than 0 and 1). Where agreement for any criterion fell below the predetermined threshold of $k \geq 0.80$, the coding scheme was discussed and clarified without reference to the sample videos, and the coders revisited their scoring. Secondary coding stopped when coders reached sufficient agreement on the sample (completion: $k = 0.92$; errors: $k = 0.88$; attempts: $k = 0.85$; serial order: $k = 0.92$) and the full data were then analyzed.

Completion time was measured in ELAN by identifying the start and end of model construction in the video: the start was defined as the moment the pieces were completely visible in the video (i.e., when the box was lifted), and the end as the moment at which the participant made the last change to the model (either attaching or removing a piece). We defined the end point based on the participant’s last action because many participants only verbally declared that they were finished after a period of examining or checking the model without actually changing it, meaning the time when they declared that they had finished often occurred many seconds (or even minutes) after they had actually finished construction.

Statistical Analysis. Performance score was calculated per participant as the sum of all four scoring criteria for all six events divided by the maximum possible score (24), which would allow for comparison across experiments. Completion time was calculated in ELAN (based on the procedure outlined above) as the time from start to finish in seconds. Both dependent variables were analyzed using a hierarchical linear regression model in R Studio (Version 1.3.959; R Core Team, 2020), with dummy-coded variables for linguistic suppression at encoding and recall (1 = *linguistic suppression*, 0 = *none*). Step 1 comprised the null (i.e., empty) model from which we extracted the Bayesian information criterion (BIC); Step 2 entered encoding and recall as fixed effects, and Step 3 entered their interaction. We used Bayesian model comparisons (BF, calculated from BIC³: Wagenmakers, 2007) to test whether the data favored the model in a given step over that of the preceding step. We also report F tests for ΔR^2 , coefficient statistics for each parameter in the best-fitting model, and estimated marginal means per condition based on the final (i.e., most complex) model.

Results

Performance in constructing the birdhouse from memory was generally good but highly variable, with scores ranging from 0.18 (poor performance, failing to construct the model) to 1.00 (perfect performance), $M = 0.63$, $SD = 0.18$. Completion time was likewise variable, ranging from 144 to 954 s ($M = 448.53$, $SD = 188.88$), and was inversely related to performance score ($r = -.63$).

Confirmatory Analyses

Performance score was affected by linguistic suppression, with evidence favoring the Step 1 model (containing fixed effects of encoding and recall) over an empty null model; (see Table 1 for model comparisons). However, there was evidence against an interaction between encoding and recall, with the data favoring the simpler Step 1 model over the Step 2 interaction model. Coefficients of the best-fitting Step 1 model indicated that performance was poorer

Table 1

Overall Fit, Change in Fit, and Model Comparisons in Hierarchical Linear Regressions of Linguistic Suppression (at Encoding and Recall) on Performance Score and Completion Time in Experiment 1

Step	Parameter(s) added	R^2	ΔR^2	F	BF ₁₀
Performance score					
1	Empty model	.000	—	—	—
2	Encoding + Recall	.156	.156	7.13	11.20
3	Encoding \times Recall (interaction)	.156	.000	0.00	0.11
Completion time					
1	Empty model	.000	—	—	—
2	Encoding + Recall	.023	.023	0.90	0.03
3	Encoding \times Recall (interaction)	.027	.004	0.35	0.13

Note. BF = Bayes factor.

when linguistic suppression was present at either encoding (unstandardized $B = -0.092$, $SE = 0.04$; standardized $\beta = -.508$, $SE = 0.208$, $t = -2.44$, $p = .017$) or recall (unstandardized $B = -0.109$, $SE = 0.04$; standardized $\beta = -.600$, $SE = 0.208$, $t = -2.88$, $p = .005$). Together, these linguistic suppression manipulations at encoding and recall explained approximately 15.6% of the variance in performance scores. Marginal means of the final model (see Figure 2) indicated performance was worst when linguistic suppression was present at both encoding and recall, and was best when linguistic suppression was not performed at all.

Analysis of completion time revealed evidence against effects of encoding and recall at Step 1 (i.e., greater evidence for the null model over the Step 1 model), and against the addition of the interaction at Step 2. Coefficients of the Step 1 model indicated that there were no significant effects at encoding (unstandardized $B = 55.79$, $SE = 42.29$; standardized $\beta = .30$, $SE = 0.224$, $t = 1.32$, $p = .191$) or recall (unstandardized $B = -9.85$, $SE = 42.29$; standardized $\beta = -.05$, $SE = 0.224$, $t = -0.23$, $p = .816$), with only 2.3% of variance explained. That is, performing linguistic suppression did not affect completion time (see Figure 2).

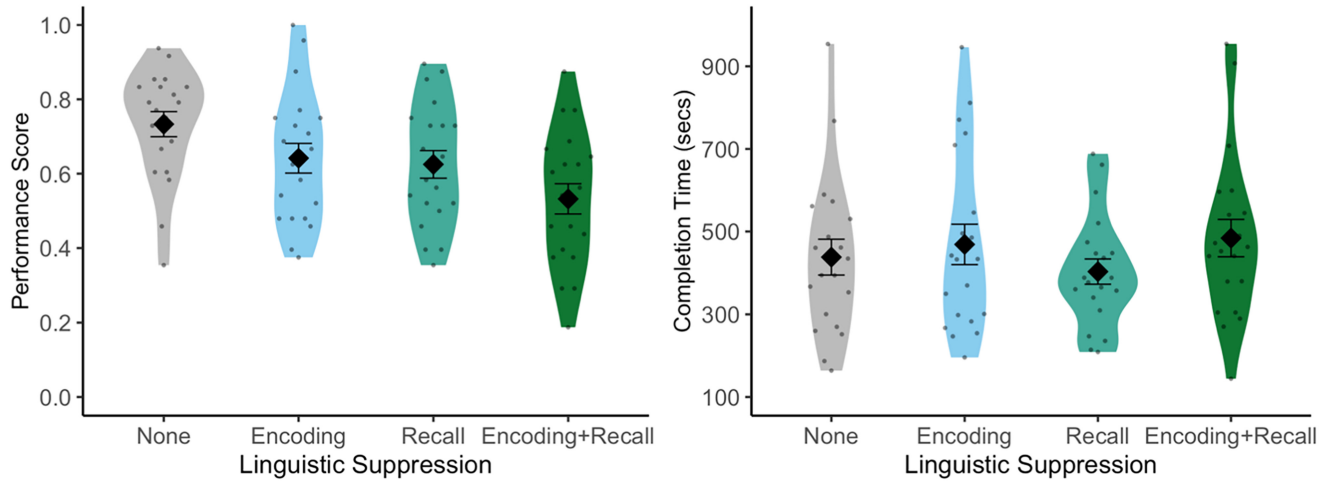
Exploratory Analyses

We estimated event memory capacity in each condition by calculating how many complete steps each participant successfully recalled and constructed in the birdhouse model. Specifically, we calculated the number of completed events (i.e., steps that scored one on the completion criterion, meaning they were fully complete in the final model), and also the more conservative number of fully accurate completed events (i.e., steps that scored one on the completion and errors criteria, meaning they were fully complete in the final model and had been constructed with no errors or deviations from instructions). These estimates of event memory capacity followed the same pattern as overall performance score. Taking the more liberal definition of successfully completed steps (which may include some errors), event memory capacity was at its best 4.7 events ($SD = 0.89$) when language was fully available (i.e., no linguistic suppression was performed during the task), but dropped to 4.0 events ($SD = 1.49$) when language was suppressed at encoding,

³ Calculating BFs using the BIC approximates a Bayesian hypothesis test assuming a unit information prior, and therefore does not require priors to be specified.

Figure 2

Performance Score and Completion Time per Condition of Performing Linguistic Suppression at Encoding and/or Recall in Experiment 1



Note. Diamonds represent means per condition with error bars of ± 1 SE. See the online article for the color version of this figure.

3.74 events ($SD = 1.41$) when language was suppressed at recall, and 3.4 events ($SD = 1.71$) when language was suppressed at both encoding and recall. Taking the more conservative definition of accurate completed events (i.e., which requires perfect reproduction from the original instructions), memory capacity was on average 3.0 events ($SD = 1.25$) when language was fully available, 2.26 events ($SD = 1.24$) when language was suppressed at encoding, 1.84 events ($SD = 1.12$) when language was suppressed at recall, and was only 1.4 events ($SD = 1.35$) when linguistic suppression was performed at both encoding and recall.

Discussion

The results of Experiment 1 partially supported our predictions. As expected, suppressing access to inner language at encoding and recall resulted in poorer performance at constructing a model from event memory and decreased memory capacity for events—participants accurately recalled 1.5 fewer events when they could not use inner language at any point during the task compared to when it was fully available throughout. These effects support the idea that inner language is routinely used to support nonverbal event memory by allowing a word or phrase to act as a placeholder in event memory and thereby bootstrapping available memory capacity. However, we did not observe the predicted interaction between linguistic suppression at encoding and recall: performance was worst when language was unavailable during encoding and recall, rather than at recall only. Furthermore, the availability of language had no effect on completion time.

These null effects led us to identify several potential methodological weaknesses in our first experiment. Firstly, the event components (i.e., the birdhouse pieces and actions to fit them together) may have been relatively difficult to label in this study, thus impeding the ability of linguistic bootstrapping to support complex event representations (i.e., reducing the impact of linguistic suppression: Nedergaard et al., 2023). Participant feedback from the debrief question “Did you use any words to help you remember?” seemed to support this interpretation, as only a limited number of labels seem to

have been consistently used as descriptors of the pieces (e.g., circle, slot, and piece) and their positions (e.g., left, beside, and opposite). Verbal descriptors produced by participants during the segmentation task also tended to be rather vague and generic (e.g., “She attaches a piece of wood to the circle”), rather than being specific enough to distinguish which piece of the birdhouse model was in question and precisely how and where it was attached. If the word labels in inner language (e.g., piece and attach) were not helpful enough to differentiate parts of the model or summarize how pieces were joined, then inner language would have limited utility in helping people to remember specific events. Secondly, the number of sub-events involved in building the birdhouse may not have strained memory capacity to sufficiently test the role of inner language, particularly since the capacity limit for recalling (and particularly reproducing) events from memory is currently unknown. The linguistic bootstrapping hypothesis predicts that language is more likely to enhance memory capacity when the sensorimotor representation in question is large or complex enough to exceed available limits, but the task only had approximately six possible events (i.e., steps) which may be well within average capacity limits (i.e., thus reducing the impact of linguistic suppression).

Thirdly, and perhaps most critically, there was only one viable end state of the birdhouse and all pieces were used in its construction, meaning that participants could potentially work out on the fly how to construct the model rather than recall the events from memory. That is, there were limited pieces and actions available to participants, so the task had limited degrees of freedom that further reduced as the task progressed (i.e., the fewer pieces that were left, the fewer the possibilities for attaching them), and participants could therefore have successfully constructed at least some model steps without relying solely on their event memory (i.e., thus reducing the impact of linguistic suppression at encoding). Moreover, given that previous studies have found that linguistic suppression negatively affects action planning (e.g., Lidstone et al., 2010; Phillips et al., 1999), we were concerned that participants could also have been using inner language at recall to support such on-the-fly planning strategies rather than retrieval from memory (i.e., thus inflating the impact of linguistic

suppression at recall but reducing its interaction with encoding). Addressing these limitations provided the impetus for the following experiments, where we moved away from the birdhouse model in favor of more complex construction models that were not subject to the same issues.

Experiment 2a: Estimating Event Memory Capacity

In this next series of experiments, we tested the linguistic bootstrapping hypothesis in an event memory task where inner language could be fully utilized by employing different models that overcame the issues of Experiment 1’s birdhouse model. The aim of the present Experiment 2a was to establish event memory capacity when inner language is fully available (i.e., without manipulating linguistic suppression), using a naturalistic event memory task similar to Experiment 1, prior to examining the role of linguistic suppression on a subset of these models in Experiment 2b.

We first developed a new set of physical models based on large, plastic construction bricks, whose components could be more easily labeled (i.e., bricks of particular shapes and/or colors), with a varying number of events involved in their construction (i.e., different models of increasing complexity), and which had multiple degrees of freedom in the reproduction task (i.e., not all available bricks were used to construct the model), thus limiting the ability to plan and construct the model on the fly. As before, participants watched an instructional video that showed the incremental construction of a particular model, and then immediately constructed the model from memory using the same components. This time, however, each participant encoded and recalled four models one after the other, where each successive model involved a longer sequence of events.

We predicted that participants’ memory performance would remain high for increasing sequence lengths until event memory capacity was reached, and would then suddenly drop once sequence length exceeded capacity. That is, longer sequences would allow more events to be recalled until the number of events exceeded memory capacity, after which point the number of recalled events would plateau out. We also expected completion time to increase with the length of the event sequence.

Method

The experiment’s design and hypotheses were preregistered at <https://aspredicted.org/y4av7.pdf>; all methods and analyses follow the preregistration unless otherwise specified. Materials, data, code, and full results output are available at <https://osf.io/v8q47/>.

Participants

A total of 20 participants took part in the study recruited via Lancaster University for payment or course credit. One participant was excluded as they failed to partially complete at least one step in each of the models. The final sample size for analysis was therefore 19 participants (15 female, $M_{\text{age}} = 23.58$ years, $SD = 5.0$). Sample size was determined via sequential hypothesis testing with BFs, where N_{min} was 16 and recruitment continued until (i) there was evidence that memory performance had begun to drop (i.e., any step model comparison exceeded the upper threshold of evidence $BF = 5$ and was stable for four successive participants); or (ii) there was evidence against any effect of sequence length on performance (i.e., all model comparisons were below the lower threshold of evidence

$BF = 0.2$ and were stable for four successive participants), or (iii) we reached the maximum sample size of $N = 32$. There was evidence for decreased memory performance at $N = 16$ for Step 2 (i.e., stopping criterion i). Due to experimenter error, an additional three participants were tested, but as this didn’t affect our findings for any of the step model comparisons (BFs were stable from $N = 16$ to $N = 19$) we have opted to report the full sample tested.

Ethics and Consent

The same two consent forms as in Experiment 1 were used for this study. Eighteen participants consented to share their video recordings (nine masked, nine with face visible) and one participant declined. All data have been shared on the OSF in accordance with participants’ individual consent choice (i.e., video data are only available for the 18 participants who opted to publicly share their videos).

Materials

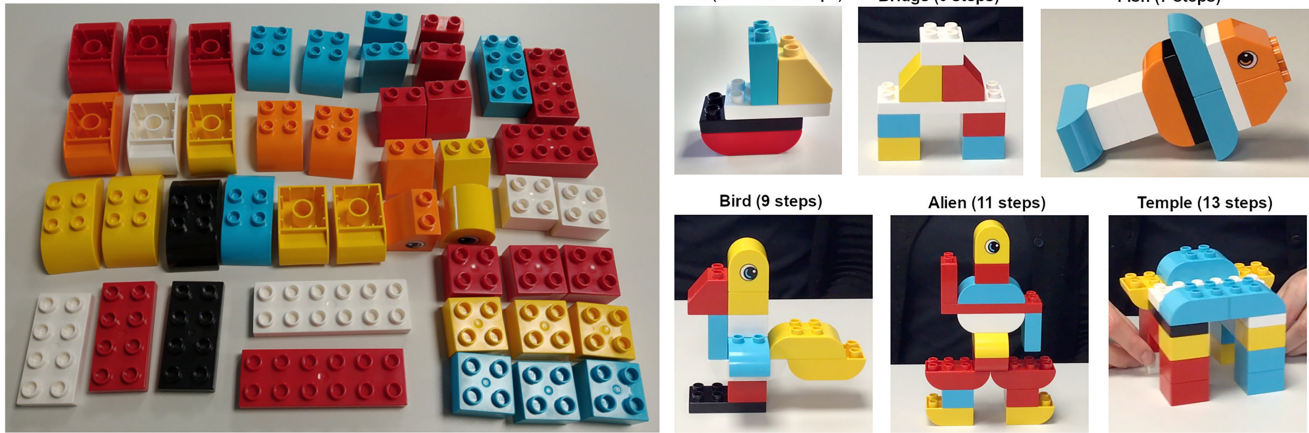
Our aim was to test five models of increasing difficulty, which varied incrementally in the number of natural events (i.e., steps) involved in constructing them. We initially developed six models (see Figure 3) using Duplo Lego bricks which are larger than standard Lego bricks and are intended for use by very young children, which ensured that the models would be easy to construct and handle, and could be clearly seen on instructional videos. Each model represented an existing object concept (a boat, a bridge, a fish, a bird, an alien, and a temple), and each comprised components that corresponded to actual parts of the object (e.g., the fish had a head, fins, and a tail). They were created using a varying number of pieces and a variety of colors and shapes. Aside from pieces with eyes painted on them, which were used as the eyes for the fish, bird and alien, the colors and shapes used were arbitrary and were not directly related to the actual concepts they represented. The models were initially piloted on four participants (who did not take part in any main experiments) to ensure that they were all constructable from memory and varied in difficulty as intended; we adjusted the models following piloting to make their increase in difficulty more incremental. In increasing number of Lego pieces involved, the final models comprised a boat (five pieces), bridge (eight pieces), fish (11 pieces), bird (12 pieces), alien (14 pieces), and temple (20 pieces). We recorded an instructional video for each model following the same procedure and using the same equipment as Experiment 1. In the videos, each model was constructed in a linear fashion, adding one piece at a time (i.e., rather than constructing subcomponents that were later joined together). Each video started with a 2-s title shown in white text on a black screen indicating the video number (practice, Video 1, Video 2, etc.), and ended with a similar 2 s title indicating “end of video.”

Segmentation Task

To establish the natural event boundaries present in constructing each model, we carried out an event segmentation task for each model following the methods and procedure for Experiment 1. An overview of the methods and results is given here, with full details provided in the additional online materials at <https://osf.io/v8q47/>. Eighteen participants ($N = 13$ female; $M_{\text{age}} = 18.67$ years, $SD = 0.59$), none of whom took part in other experiments, watched one practice video from Experiment 1, followed by the six instructional

Figure 3

Array of Lego Pieces Presented to Participants (Including Distractor Pieces) and the Six Models in Experiment 2a



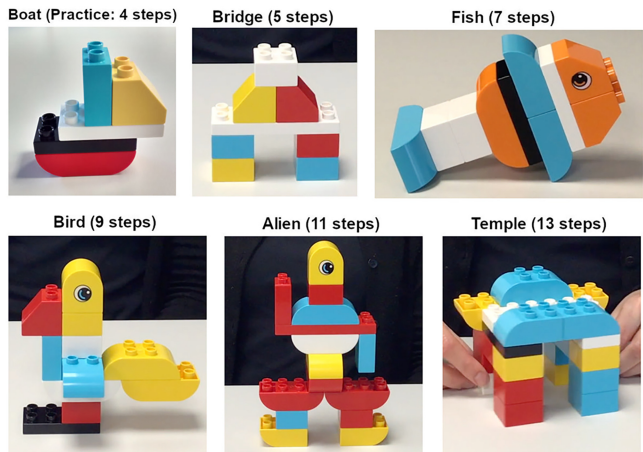
Note. See the online article for the color version of this figure.

videos created for the Lego models. The videos were always played in the same order of increasing number of pieces, starting with the simplest model. The segmentation task instructions, procedure, data preparation, and analysis (i.e., to identify meaningful event boundaries) were identical to Experiment 1.

Participants showed overall high agreement for identifying boundaries of critical events, with mean agreement $>80\%$ for each model. The number of events present in the models increased incrementally with, but did not directly correspond to, the number of Lego pieces involved ($r = .970$), resulting in the following numbers of events per model: boat = 4, bridge = 5, fish = 7, bird = 9, alien = 11, and temple = 13. We therefore considered these events to be the necessary steps to complete each model, and used them as the basis for coding performance in this and subsequent experiments. We decided to use the boat model as a practice item in the main experiment, while the remaining five models would be the test items; thus, our models tested memory capacity for sequences of between five and 13 events.

Design and Procedure

We used a within-participant design where all participants watched all videos and constructed all models under the same conditions and in the same order. They first watched the video for the boat model and constructed it from memory as a practice item, followed by each video and corresponding model construction in increasing levels of difficulty: bridge (five steps), fish (seven steps), bird (nine steps), alien (11 steps), and temple (13 steps). At no point were participants given descriptive labels for any of the models (i.e., boat, bridge, fish, etc.); we use these names here purely for reporting purposes. After watching each video, participants sat at a different table and were presented with an array of Duplo Lego pieces including distractors; all pieces in the array were displayed in exactly the same configuration for each participant, and for each model (see Figure 3), and were covered until the participant was ready to start. Distractors were selected based on the following two criteria: (a) For every piece used in a model, there was at least one valid distractor piece based on its color, shape or other features (e.g., if a black piece was used in any



of the models, at least one other black piece was present in the array; or if a long flat piece was used in any of the models, at least one other piece with exactly the same shape was present in the array); (b) Where two pieces of the same color or shape were used in a model, two other pieces of the same color or shape were also present in the array (e.g., if two orange pieces were used in the model, there were at least two other orange pieces present in the array). Some of the distractors were used in other models (e.g., the distractor pieces for the bird might be used in the fish or alien); thus, the array comprised 45 pieces in total, and included nine distractors that were not used in any of the models (see Figure 3).

Other than these differences, the procedure was identical to Experiment 1, except that participants were not trained in or asked to perform linguistic suppression, and each instructional video was played via QuickTime Player. Immediately before constructing the practice model, participants were reminded that they should construct the model exactly as they had seen in the video, and that there was no time limit but they should let the experimenter know when they had finished or when they could not get any further. Participants were not given any further instructions or tasks immediately before constructing the five test models, except being told when they could start.

Data Preparation and Analysis

Video Coding and Scoring. Prior to testing, we updated the scoring criteria from Experiment 1 (completion, errors, multiple attempts, and serial order) for the new types of model, particularly regarding how to identify which steps participants intended to construct when they looked very different to the instructions (e.g., when participants incorrectly used using distractor pieces), and to clarify the Errors criterion based on the color, shape, and position of pieces. Based on the events identified and labeled in the segmentation task, we created clear definitions of each model step, what pieces it should include to be scored correctly, and what overall location in the model each step related to (e.g., at the top, middle, or bottom; at the front or back); these definitions, along with the scoring criteria, are available at <https://osf.io/v8q47/>.

Participant videos were again coded in ELAN software and scored according to the updated criteria following exactly the same procedure as Experiment 1. A second coder, blind to the aims and conditions of the study, scored a sample of 20 videos (21% of all videos, pseudo-randomly selected cover a systematic cross-section of sequence lengths and summed scores), to ensure that the scoring criteria could be objectively and consistently applied. Following training on coding videos of all five models (not included in the sample), the second coder independently scored the sample videos and compared to the first coder's using the same method as in Experiment 1. Where agreement for any criterion fell below the predetermined threshold of Cohen's $k \geq .80$, the coding scheme was discussed and clarified without reference to the sample videos, and the coders revisited their scoring. Secondary coding stopped when coders reached sufficient agreement on the sample (completion: $k = 0.91$; errors: $k = 0.89$; multiple attempts: $k = 0.84$; serial order: $k = 0.83$) and the full data were analyzed.

Statistical Analysis. Participant performance was calculated as a ratio score (i.e., the sum of all four scoring criteria divided by the maximum possible score for the model) which allowed comparison across models of varying sequence length. Completion time was calculated in ELAN as the time from start to finish in seconds.⁴ These measures were analyzed using hierarchical linear mixed models with participant as a random effect and sequence length as fixed effects, using the lme4 package (Version 1.1-23; Bates et al., 2015) in RStudio (Version 1.3.939; R Core Team, 2020). Sequence length was reverse Helmert-coded to compare the effect of each sequence length with the mean of the previous (shorter) sequences, which produced four coded variables: seven versus five steps (fish vs. bridge), nine versus five to seven steps (bird vs. bridge and fish), 11 versus five to nine steps (alien vs. bridge, fish, and bird), and 13 versus five to 11 steps (temple vs. bridge, fish, bird, and alien). This coding method is suited to capturing nonlinear monotonic trends (e.g., a plateau followed by a fall, or vice versa) and allowed us to determine the tipping point at which performance score dropped due to the sequence length surpassing event memory capacity. In the hierarchical model, Step 0 (null model) included participant as random effect, Step 1 added sequence length as seven versus five steps, Step 2 added sequence length as nine versus five to seven steps, Step 3 added sequence length as 11 versus five to nine steps, and Step 4 added sequence length as 13 versus five to 11 steps. We ran Bayesian model comparisons between successive regression steps as per Experiment 1, where the first hierarchical step with evidence in its favor represented the sequence length at which performance score differed from that of shorter sequences (i.e., the point at which performance started to decline because event sequence length exceeded event memory capacity). This point of inflection was used to estimate event memory capacity. We also report marginal R^2 (calculated using the MuMIn package Version 1.43.17; Barton, 2017), coefficient statistics for each parameter in the final (Step 4) model, and estimated marginal means per condition based on the final model.

Results and Discussion

All reported analyses are confirmatory. Performance in constructing each model from memory was overall good and, as expected, event memory performance generally increased as sequence length increased (see Figure 4). In performance score, Bayesian model comparisons did not favor any effect of sequence length while the

sequence comprised seven events or fewer: Step 1 did not improve model fit over Step 0 (i.e., data equivocally favored the null Step 0 at $BF_{01} = 4.38$). However, there was very strong evidence for Step 2 over Step 1 ($BF_{10} = 3,542.97$), indicating that the data favored a model that distinguished sequences of nine events from shorter sequences. Likewise, there was strong evidence for Step 3 over Step 2 ($BF_{10} = 1,168.51$), and for Step 4 over Step 3 ($BF_{10} = 292.78$). That is, participants could recall sequences of five or seven steps with comparable performance, suggesting that seven events are at or within capacity limits of event memory. However, their performance dropped markedly for sequences of nine steps and continued to drop for sequences of 11 and 13 steps as a smaller proportion of events could be recalled, meaning that sequences of nine steps or more exceeded capacity limits. The final (Step 4) model of sequence length explained a large proportion of variance in performance score ($R^2 = .392$); coefficient statistics per Helmert-coded parameter can be seen in Table 2.

In completion time, people also generally took longer to complete the models as sequence length increased (see Figure 4). Bayesian model comparisons found no evidence of difference in completion times for sequences of five and seven steps (Step 1 did not positively improve model fit over the Step 0 null model, equivocal $BF_{10} = 0.28$). However, performance slowed down for sequences of nine steps (strong evidence for Step 2 over Step 1, $BF_{10} = 25.41$), and again at 11 and 13 steps (very strong evidence for Step 3 over Step 2, $BF_{10} = 54,727.07$; and for Step 4 over Step 3, $BF_{10} = 37,304,355$). The final (Step 4) model of sequence length explained almost half the variance in completion times ($R^2 = .490$); coefficient statistics are reported in Table 2.

Together, these results suggest that the maximum memory capacity for events is approximately seven to eight events when language is fully available, which in turn suggests that the birdhouse task in Experiment 1 (with six events) may indeed have been subject to ceiling effects. Short sequences of up to seven (or potentially eight) events can be remembered and reproduced relatively quickly and accurately, but increasingly longer sequences of nine events or more are distinguished by increasingly poorer, slower performance.

Experiment 2b: Does Access to Inner Language Enhance Memory More for Longer Sequences of Events?

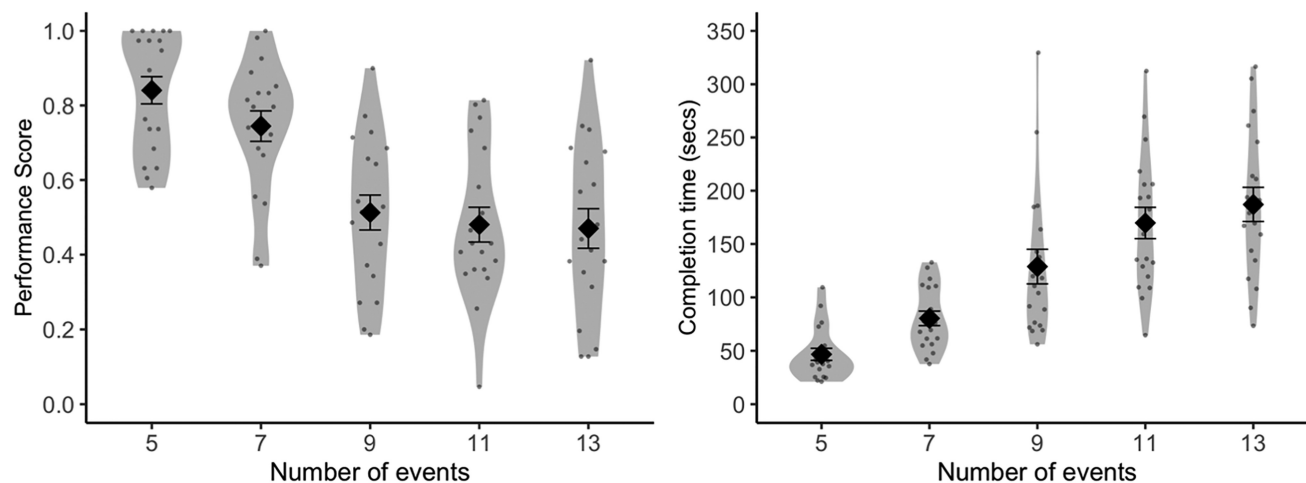
The aim of Experiment 2b was to retest the linguistic bootstrapping hypothesis in sequences of events that were either within, or which clearly exceeded, event memory capacity. Results from Experiment 2a had estimated event memory capacity to be between seven and eight events, so we selected the models with seven and 11 steps (the fish and alien from Figure 3) to use in the present experiment. Each participant constructed both models from memory, starting with the shorter sequence. As in Experiment 1, access to inner language was manipulated by asking participants to perform linguistic suppression at either encoding or recall.

We first predicted that memory performance would be poorer, and completion time longer, for the longer 11-event sequence. Secondly, we predicted that memory performance and completion time would be impaired when access to linguistic labels was limited during

⁴ The preregistration stated minutes due to an oversight, as ELAN software records timings in seconds.

Figure 4

Performance Score and Completion Time for Each Sequence Length (Model) in Experiment 2a



Note. Diamonds represent means per condition with error bars of $\pm 1 SE$.

encoding and/or recall using linguistic suppression. Lastly, we predicted that reliance on linguistic labels would be greater for the longer sequence of 11 events that exceeded memory capacity compared to the shorter sequence that was within capacity; that is, we predicted an interaction between the availability of inner language and sequence length whereby linguistic suppression would have a greater effect on performance in the longer sequence compared to the shorter sequence. We predicted that memory performance would be best and completion time fastest for the shorter sequence length when inner language was available at both encoding and recall, while performance would be poorest and completion time slowest for the longer sequence length when inner language was unavailable at encoding and recall.

Method

The experiment's design and hypotheses were preregistered at <https://aspredicted.org/gc3yf.pdf>; all methods and analyses follow the preregistration unless otherwise specified. Materials, data, code, and full results output are available at <https://osf.io/v8q47/>.

Table 2

Unstandardized Regression Coefficients and Associated Statistics for Sequence Length (Helmert-Coded) in the Final Step 4 Regression Model of Performance Score and Completion Time in Experiment 2a

Parameter(s) added	Coefficient	SE	<i>t</i>	<i>p</i>
Performance score				
7 versus 5 steps	-0.096	0.052	-1.831	.067
9 versus 5–7 steps	-0.280	0.045	-6.158	<.001
11 versus 5–9 steps	-0.219	0.043	-5.110	<.001
13 versus 5–11 steps	-0.175	0.042	-4.207	<.001
Completion time				
7 versus 5 steps	33.625	14.201	2.368	.018
9 versus 5–7 steps	65.304	12.298	5.310	<.001
11 versus 5–9 steps	84.383	11.595	7.278	<.001
13 versus 5–11 steps	80.699	11.227	7.188	<.001

Participants

A total of 74 participants recruited via Lancaster University took part for payment or course credit. Five were excluded based on pre-registered inclusion criteria (four were nonnative speakers of English and one participant failed to partially complete at least one step in each of the models). A further five participants were excluded for additional reasons (one had taken part in Experiment 1; three did not carry out linguistic suppression correctly or additionally tapped their foot; and one due to technical issues with video playback). The final sample size for analysis was 64 participants (48 female; $M_{\text{age}} = 21.13$ years, $SD = 6.73$). Sample size was determined via sequential hypothesis testing with BFs, where N_{min} was set at 64 (eight participants per condition) and recruitment continued until the interaction between linguistic suppression and sequence length cleared the prespecified grade of evidence $BF \geq 5$ for memory performance score (i.e., there was evidence for or against the best-fitting Step 4 model over the Step 3 model—see Statistical Analysis for details), and was stable for four successive participants, or when we reached the N_{max} of 128 (16 participants per condition). We found stable evidence against an interaction for all Step 4 models at $N = 64$ (i.e., $BF_{10} < 0.2$) and so recruitment stopped at N_{min} .

Ethics and Consent

The same two consent forms as in Experiment 1 were used for this study. Fifty-six participants consented to share their video recordings (25 masked, 31 with face visible) and eight participants declined. All data have been shared on the OSF in accordance with participants' individual consent choice (i.e., video data are only available for the 56 participants who opted to publicly share their videos).

Materials

As test items, we used two models from Experiment 2a whose sequence lengths were identified as being within and beyond event

memory capacity: the fish (seven steps) and the alien (11 steps). As before, the boat model (four steps) was used as a practice item. The array of Lego pieces presented to participants again included distractors using the same selection process as in Experiment 2a. The array comprised 31 pieces, eight of which were not included in either of the models, and was presented in the same configuration for every participant. Instructional videos were exactly as per Experiment 2a.

Design and Procedure

Sequence length and the availability of inner language (i.e., using linguistic suppression as a secondary task) were manipulated within participants, where the timing of linguistic suppression (i.e., at encoding or recall) was manipulated across participants. That is, participants completed both models in order of increasing sequence length, and each model was completed under one of the four possible combinations of linguistic suppression conditions (none at all, encoding, recall, or encoding + recall). By rotating linguistic suppression conditions across models, we ensured that no participant experienced the same conditions for both sequence lengths (i.e., the experience of constructing the second model was always different from the first), which produced a full $2 \times 2 \times 2$ design across the experiment as a whole: Sequence Length (Seven Steps, 11 Steps) \times Encoding (Suppression, No Suppression) \times Recall (Suppression, No Suppression).

The procedure was the same as in Experiment 1 except that all participants performed linguistic suppression for at least one of the test models (at either encoding, recall, or both), and so all participants were given instructions and an opportunity to practice linguistic suppression before starting the experiment. Specifically, participants were told that, at some point during the study, they were going to be asked to say “the” out loud repeatedly at a rate of approximately two words per second, and that the experimenter would tell them exactly when to start and stop; no participant knew in advance when they were going to be asked to do this. Prior to starting the experiment, participants practiced linguistic suppression for a few seconds with the experimenter using a metronome app to achieve the correct rate. All participants then completed the event memory task in the same order: they first watched the instructional video and constructed the practice model, followed by the seven-step fish model and then the 11-step alien model. Experimental setup and procedure were otherwise identical to Experiment 2a.

Data Preparation and Analysis

Video Coding and Scoring. Participant videos were coded in ELAN software and scored according to the final coding scheme for Experiment 2a, and following exactly the same procedure. The second coder from Experiment 2a (still blind to the aims and conditions of the study) independently coded and scored a sample of 24 videos (19% of all videos, selected pseudo-randomly to cover a systematic cross-section of conditions and summed scores), to ensure that the scoring criteria could be objectively and consistently applied. The second coder independently scored the sample videos which were then compared to the first coder’s using the same method as in Experiment 1. Where agreement for any criterion fell below the predetermined threshold of Cohen’s $k \geq .80$, the coding scheme was discussed and clarified without reference to the sample videos, and

the coders revisited their scoring. Secondary coding stopped when coders reached sufficient agreement on the sample (completion: $k = 0.87$; errors: $k = 0.84$; multiple attempts: $k = 0.81$; serial order: $k = 0.81$). All remaining videos were then recoded and scored by the primary researcher following the modified scoring criteria, and the full data were analyzed.

Statistical Analysis. Participant performance score and completion time were measured in the same way as in Experiment 2a. Three trials were excluded from the completion time analysis (two for the seven-step fish model, one for the 11-step alien model) as completion times were >3 SDs from the mean for that model. Dependent variables were analyzed using hierarchical linear mixed-effects models using the lme4 package (Version 1.1-23; Bates et al., 2015) in RStudio (Version 1.3.959; R Core Team, 2020), with participant as a random effect and fixed effects of linguistic suppression and sequence length. Linguistic suppression at encoding and recall was dummy-coded (1 = *linguistic suppression*, 0 = *none*), and sequence length was entered as a categorical variable (default dummy-coded with reference level of longer sequence; i.e., the 11-step alien model). Step 0 comprised the baseline model of participant as a random effect; Step 1 entered sequence length, Step 2 entered the timing of linguistic suppression (encoding and recall), and Step 3 entered the Encoding \times Recall interaction. Step 4 examined candidate interactions a–d between sequence length and the timing of linguistic suppression (see Table 3 for all model parameters per step).

We compared successive hierarchical steps using BFs (calculated via BIC; Wagenmakers, 2007). The Step 1 model comparison (i.e., against preceding Step 0) tested the hypothesis that performance score would be lower and completion time longer for the longer sequence, and the Step 2 model comparison tested the hypothesis that performance score would be lower and completion time longer with linguistic suppression at encoding and recall. To test whether there was an interaction between sequence length and linguistic suppression, we compared each Step 4 model from candidates a–d (see Table 3) in turn to Step 3 and selected the one with strongest evidence (i.e., the largest BF); we then used that best-fitting Step 4 model as the basis of our inference regarding the interactions. We also report marginal R^2 (calculated in R using the MuMIn package Version 1.43.17; Barton, 2017) per regression step, coefficient statistics for the best-fitting model, and estimated marginal means per condition based on the final (i.e., most complex) model.

Results

Performance in constructing both models was overall good but highly variable (see Figures 5 and 6). Performance scores ranged from 0.20 to 0.96 ($M = 0.58$, $SD = 0.21$) for the seven-step fish model, and from 0.02 to 0.86 ($M = 0.41$, $SD = 0.18$) for the 11-step alien model. Completion times ranged from 34 to 228 s ($M = 89.73$, $SD = 44.43$) for the seven-step fish model, and from 67 to 465 s ($M = 181.87$, $SD = 95.17$) for the 11-step alien model, and were inversely related to performance scores ($r = -.578$).

Confirmatory Analyses

Performance scores were affected by sequence length and linguistic suppression, with strong evidence favoring the Step 1 model (containing the fixed effect of sequence length) over the Step 0

Table 3

Overall Fit, Change in Fit, and Model Comparisons in Hierarchical Linear Regressions of Sequence Length and Linguistic Suppression (at Encoding and Recall) on Performance Score and Completion Time in Experiment 2b

Step	Parameter(s) added	R^2	ΔR^2	BF ₁₀
Performance score				
0	Participant as random effect	.000	—	—
1	Sequence length	.165	.165	14,395.75
2	Encoding + Recall	.286	.121	195.84
3	Encoding × Recall	.296	.010	0.25
4a	Sequence length × Encoding	.300	.004	0.13
4b	Sequence length × Recall	.297	.000	0.09
4c	Sequence length × Encoding + Sequence length × Recall	.300	.004	0.01
4d	Sequence length × Encoding + Sequence length × Recall + Sequence length × Encoding × Recall	.300	.004	0.00
Completion time				
0	Participant as random effect	.000	—	—
1	Sequence length	.273	.273	568,299,949
2	Encoding + Recall	.306	.033	0.20
3	Encoding × Recall	.306	.033	0.09
4a	Sequence length × Encoding	.306	.033	0.09
4b	Sequence length × Recall	.306	.034	0.09
4c	Sequence length × Encoding + Sequence length × Recall	.306	.034	0.01
4d	Sequence length × Encoding, Sequence length × Recall + Sequence length × Encoding × Recall	.307	.034	0.00

Note. BF = Bayes factor.

model (with participant as random effect),⁵ and favoring the Step 2 model (containing fixed effects of encoding and recall) over Step 1; see Table 3 for model comparisons. However, there was equivocal evidence against an interaction between the timing of linguistic suppression at encoding and recall (i.e., evidence equivocally favored the Step 2 model over Step 3), and clear evidence against any interactions between sequence length and performance of linguistic suppression at either time point (i.e., evidence favored the Step 3 model over the best of the Step 4 candidates, Model 4a). Coefficients for the best-fitting Step 2 model indicated that performance was better for the shorter seven-step model (unstandardized $B = 0.171$, $SE = 0.03$; standardized $\beta = .810$, $SE = 0.14$, $t = 5.85$, $p < .001$) than the longer 11-step model, but was overall worse across both models when linguistic suppression was performed at encoding (unstandardized $B = -0.146$, $SE = 0.03$; standardized $\beta = -.690$, $SE = 0.15$, $t = -4.68$, $p < .001$). However, when linguistic suppression was performed at recall, it did not affect performance (unstandardized $B = 0.005$, $SE = 0.03$; standardized $\beta = .022$, $SE = 0.15$, $t = 0.15$, $p = .880$). That is, manipulating linguistic suppression explained 12.1% of the variance in performance scores above and beyond the effect of sequence length, and the vast majority of this linguistic suppression effect was due to its manipulation at encoding rather than recall. Marginal means of Model 4d (see Figure 6) indicated performance was worst when linguistic suppression was present at encoding in the 11-step model, and was best in the seven-step model when linguistic suppression was either not performed at all or only performed at recall.

Completion time was affected by sequence length, with strong evidence favoring the Step 1 model (containing the fixed effect of sequence length) over the Step 0 model of participant as a random effect. However, there was evidence against an effect of linguistic suppression on completion time (evidence favored the Step 1 model over Step 2), and against any interactions between the timing of linguistic suppression at encoding and recall (i.e., evidence again favored Step 2 over Step 3) or between sequence length

and linguistic suppression at either time point (i.e., evidence again favored Step 3 over the best Step 4 candidate, Model 4b). Figure 6 shows marginal means per condition from Model 4d. Coefficients from the best-fitting Step 1 model indicated that participants were quicker to complete the shorter seven-step model (unstandardized $B = -90.402$, $SE = 11.32$; standardized $\beta = -.940$, $SE = 0.12$, $t = -7.99$, $p \leq .001$) than the longer 11-step model.

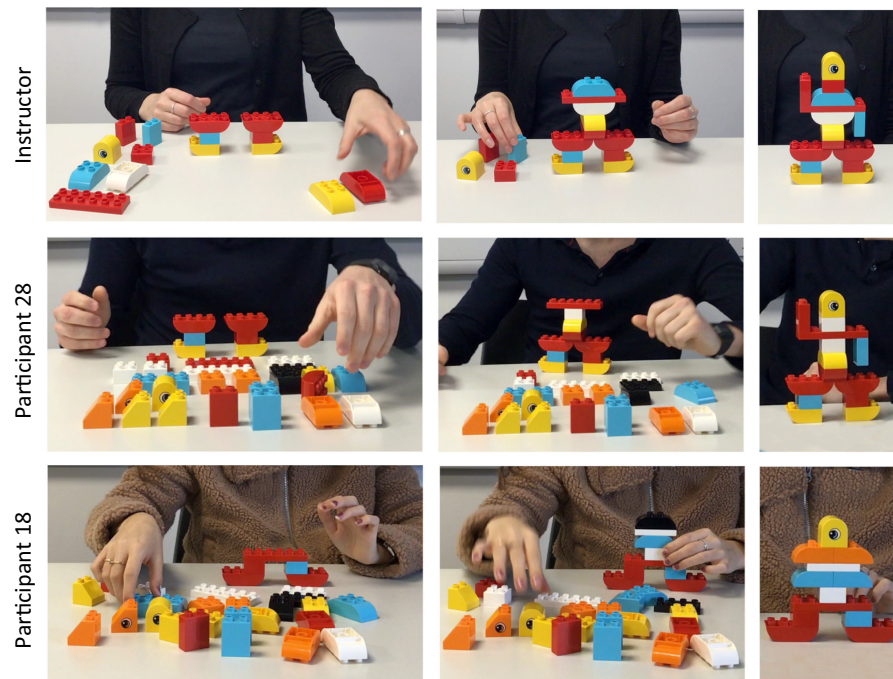
Exploratory Analyses

As in Experiment 1, we estimated event memory capacity in each condition by calculating how many complete steps each participant successfully recalled and constructed in each model. We again calculated the number of completed events (i.e., steps that scored 1 on the Completion criterion, meaning they were fully complete in the final model), and the more conservative number of fully accurate completed events (i.e., steps that scored 1 on the completion and errors criteria, meaning they were fully complete in the final model and had been constructed with no errors or deviations from instructions). As in Experiment 1, these estimates of event memory capacity followed a similar pattern to overall performance score; see Table 4 for descriptive statistics. For the more liberal measure of completed events, memory capacity was a little over four events when inner language was available during encoding (i.e., when linguistic suppression was performed at recall only or not at all), irrespective of whether the original sequence comprised seven or 11 events. However, when linguistic suppression was performed at encoding, this capacity estimate dropped to approximately three events for the seven-step model and two events for the 11-step

⁵ The Step 0 model (random effects only) produced a singular fit in analysis of both performance score and completion time, which may make model comparisons with Step 1 unreliable. We, therefore, explored an alternative model without random effects (e.g., Barr et al., 2013), which produced the same pattern of effects as the original analysis (see the additional online materials for details), and so we opted to retain the pre-registered analysis.

Figure 5

Early, Middle, and Final Stages of Construction of the 11-Step Alien Model in Experiment 2b, Taken From the Instructional Video (Top Row), a High-Performing Participant (Middle Row), and a Poor-Performing Participant (Bottom Row)



Note. Full videos can be viewed on OSF; participant numbers relate to Experiment 2b only. OSF = Open Science Framework. See the online article for the color version of this figure.

model. Taking the more conservative measure of fully accurate completed events, memory capacity was between three and four events when language was unavailable at encoding, and dropped to one to two events when language was unavailable at recall. Regardless of whether capacity for complete events was estimated liberally or conservatively, disrupting access to inner language led to approximately one to two fewer events being recalled. Compared to the baseline memory capacity of seven to eight events determined in Experiment 2a, these capacity estimates are much lower because they focus on complete events rather than partial representations of multiple events. We return to this point in the General Discussion section.

Discussion

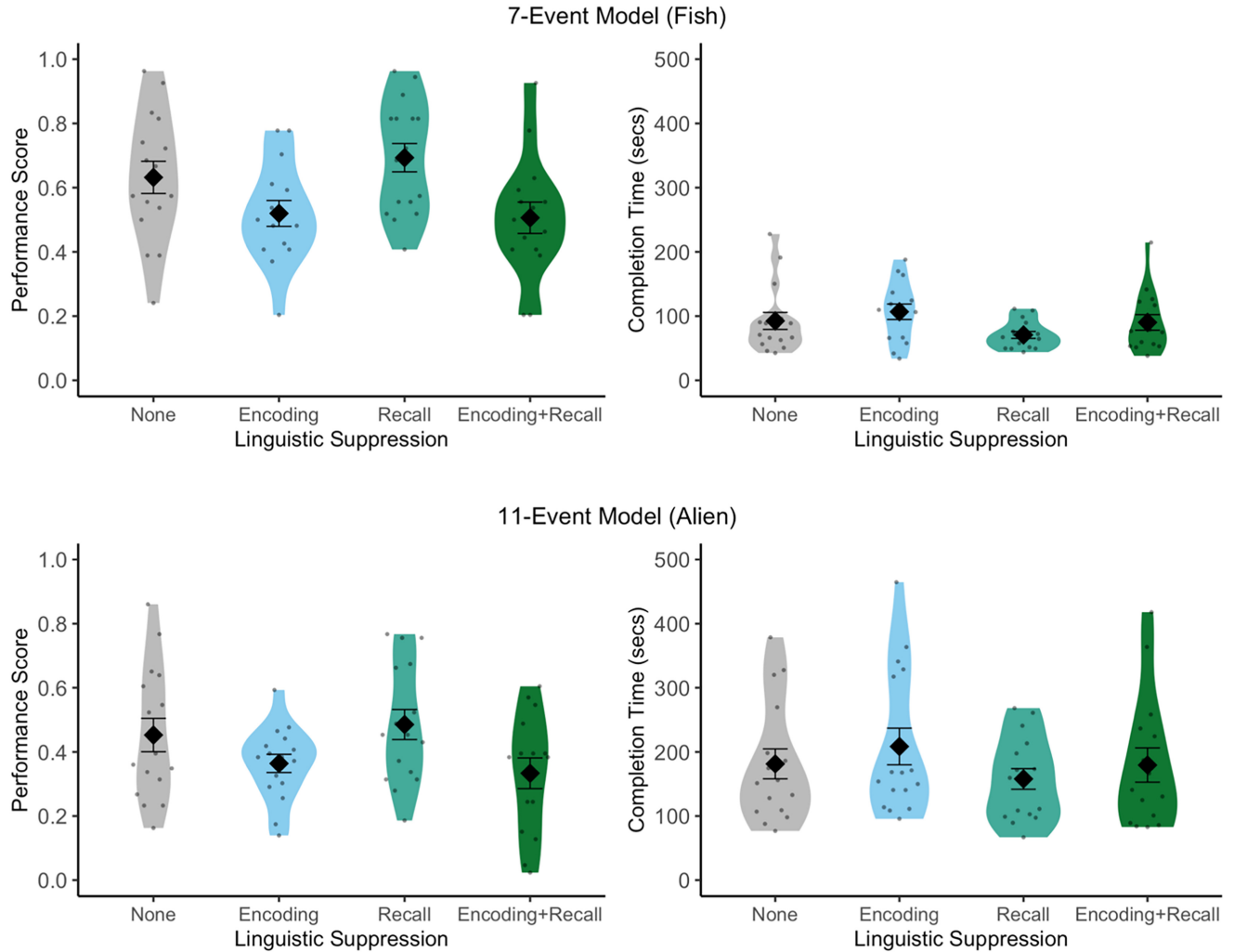
Experiment 2b addressed several limitations of Experiment 1: we used different models that could be more easily labeled, had more degrees of freedom to prevent participants planning their constructions on the fly, had a larger number of subevents (steps) to avoid ceiling effects, and we compared all conditions within participants to avoid group differences. We also compared different sequence lengths that were within and which exceeded memory capacity, which allowed us to test whether the use of inner language was more likely in the longer sequence. With these methodological improvements, results partially supported our hypotheses. As predicted, memory performance was overall worse for longer sequences, consistent with Experiment 2a. Also as predicted, and

replicating our findings from Experiment 1, suppressing access to inner language during encoding resulted in poorer memory performance, which supports the idea that inner language is indeed used to encode events in memory. Disruption of inner language at recall had no effect on memory performance in this experiment, meaning that (as earlier speculated) the observed effect in Experiment 1 could have been due to participants using inner language to plan their actions on the fly when constructing the models rather than to support representing events in memory. However, contrary to our predictions, there was no interaction between sequence length and linguistic suppression, suggesting that inner language is equally beneficial to memory for events regardless of whether memory capacity is strained or not. We return to these null findings in the general discussion. Finally, as in Experiment 1, though again against our predictions, linguistic suppression did not affect completion time nor did it interact in any way with sequence length. These findings suggest that while inner language may benefit event memory by allowing people to remember a larger number of events, it does not actually help people to recall those events any more rapidly.

Both Experiments 1 and 2b have shown that linguistic suppression at encoding consistently affected memory for events in; however, as linguistic suppression was the only secondary task employed, the possibility remains open that it could merely be a dual-task effect (i.e., due to the extra demands of performing a secondary task). We therefore conducted a final experiment using a control task to establish whether the observed effects were specifically due to disruption of inner language.

Figure 6

Performance Score and Completion Time for the Seven-Step and 11-Step Models in Experiment 2b per Condition of Performing Linguistic Suppression at Encoding and/or Recall



Note. Diamonds represent means per condition with error bars of $\pm 1 SE$. OSF = Open Science Framework. See the online article for the color version of this figure.

Experiment 3: Dual-Task Control Study

In this final experiment, we tested whether the results of Experiments 1 and 2b were indeed due to suppression of inner language, or whether they were simply due to dual-task interference caused by the extra demands of carrying out a concurrent secondary task. We used the same event reproduction task as in Experiment 2b, this time selecting three models of differing sequence length. Critically, as well as linguistic suppression, we also asked participants to perform an auditory monitoring task (i.e., monitor a sequence of rhythmic clicks and respond when a beep sounds), which provided a control secondary task of equivalent difficulty that left inner language fully available. By comparing memory performance between the two secondary tasks, we could therefore separate effects of dual-task interference from that of disrupting access to inner language. We manipulated the secondary tasks at encoding

only, as linguistic suppression had consistently affected event memory during encoding in Experiments 1 and 2b (although an effect during recall was found in Experiment 1, this was not replicated in Experiment 2b). Finally, we also examined the effects on memory performance score only as linguistic suppression had had no effect on completion time in the previous experiments.

We hypothesized that memory performance would be poorer when inner language was unavailable during encoding (i.e., when linguistic suppression was performed) compared to when language was available (i.e., with no secondary task), replicating the findings from Experiments 1 and 2b. Importantly, we also predicted poorer memory performance when participants performed linguistic suppression compared to the auditory monitoring control task; that is, we expected performance to be best when participants performed no secondary task, and worst with linguistic suppression. We also predicted, as per Experiments 2a and b, that memory performance

Table 4
Estimated Means (Standard Deviations) of Event Memory Capacity for the Seven-Step and 11-Step Models in Experiment 2b

Sequence length	Linguistic suppression			
	None	Encoding	Recall	Encoding + recall
Completed events				
7 steps	4.13 (1.93)	3.44 (2.10)	4.50 (1.83)	2.69 (1.78)
11 steps	4.19 (2.99)	2.50 (1.86)	4.81 (3.17)	2.00 (2.10)
Fully accurate completed events				
7 steps	3.19 (1.97)	2.06 (1.53)	4.06 (1.88)	1.81 (1.72)
11 steps	2.94 (2.70)	1.50 (1.97)	3.50 (3.01)	1.38 (1.50)

would be poorer for longer sequences of actions compared to shorter.

Method

The experiment’s design and hypotheses were preregistered at <https://aspredicted.org/hf2gu.pdf>; all methods and analyses follow the preregistration unless otherwise specified. Materials, data, code, and full results output are available at <https://osf.io/v8q47/>.

Participants

A total of 34 participants took part in the study recruited via Lancaster University for payment or course credit, but seven were excluded: four because they were nonnative speakers of English, one performed the auditory monitoring control task incorrectly, one was wearing a face mask during linguistic suppression,⁶ and one experienced technical issues. The final sample size for analysis was 27 (19 female; $M_{\text{age}} = 18.6$ years, $SD = 1.24$). Sample size was determined via sequential hypothesis testing with BFs, where N_{min} was set at 21 (seven participants per condition). Recruitment continued until the difference in memory performance between linguistic suppression and the auditory monitoring task cleared the prespecified grade of evidence $BF \geq 5$ or its reciprocal $1/5$ (i.e., there was evidence for or against the Step 3 over the Step 2 model—see Statistical Analysis section) and was stable for three successive participants, or until we reached the N_{max} of 42 (14 participants per condition).⁷ We found stable evidence for a difference at $N = 27$ ($BF_{10} = 47.56$; full results available at <https://osf.io/v8q47/>) and so stopped recruitment.

Ethics and Consent

The same two consent forms as in Experiment 1 were used for this study. Twenty-five participants consented to share their video recordings (nine masked, 16 with face visible) and two participants declined. All data have been shared on the OSF in accordance with participants’ individual consent choice (i.e., video data are only available for the 25 participants who opted to publicly share their videos).

Materials

We used three of the Lego models (and corresponding instructional videos) from Experiment 2a: the fish (seven steps), the bird (nine steps), and the alien (11 steps), as well as the boat model

(four steps) as a practice item. The array of Lego pieces presented to participants again included distractor pieces that were selected using the same process as in Experiments 2a and 2b. The array comprised 35 pieces, eight of which were not included in any of the models, and was presented in the same configuration for every participant.

We created sound files for the auditory monitoring task using Audacity software (Version 3.0.4; Audacity Team, 2019). The sound files comprised a continuous series of auditory clicks at the rate of two per second, interspersed with two pure tones (beeps) at particular intervals that were matched to each instructional video. The background click sound was a “finger snap” percussion loop sourced from a website of free music samples (MusicRadar, n.d.). It was sampled at a repeating interval of 500 ms between click onsets, and used to create a background soundtrack with the same duration as each instructional video. The pure tones were set at 440 Hz and lasted 500 ms. We then created nine sound files for each instructional video by inserting the tones at different intervals in the background soundtrack in order to jitter the time point during model instruction that participants would hear the tones. Tones were always inserted “offbeat” to the background clicks (i.e., they never started playing at the same time as a click). The timing of the tones occurred within a fixed time window: tones were never played during the first or last 10 s of a given video’s duration, or in the middle 15 s. As each instructional video (and its accompanying sound files) lasted a different length, it meant that the time windows for playing the tone were slightly different for each video: for the seven-step fish model, tones were played between 10–19.5 s and 34.5–44 s; for the nine-step bird model, they were played between 10–25 and 40–55 s; and for the 11-step alien model, they were played between 10–32 and 47–69 s). The sound files were then exported at a sample rate of 44,100 Hz as digital .wav files and attached to the corresponding instructional videos using iMovie Version 10.2.5 (all videos are available at <https://osf.io/v8q47/>).

Auditory Monitoring Control Task

For a secondary task to act as a suitable control for linguistic suppression, it should match its demands as closely as possible without affecting any linguistic processes (i.e., leaving access to inner language intact). Within these constraints, we judged many common secondary tasks to be unsuitable for our purposes. For instance, although finger or foot tapping is a common control task in studies using linguistic suppression, we did not deem it suitable for the present experiment because previous research has shown that performing repetitive actions interferes with memory for bodily actions (Smyth & Pendleton, 1989; see also Shebani & Pulvermüller, 2013), and such repetitive action tasks may thus have interfered with how participants encoded the sensorimotor information involved in constructing the models. Moreover, foot tapping lacked

⁶As the study was conducted during the COVID-19 pandemic, participants were required to wear facemasks on entering the lab; due to experimenter error, one participant was not asked to remove their face mask before performing linguistic suppression.

⁷We were able to specify smaller N_{min} and N_{max} sample size than in Experiments 1 and 2b due to the experimental design (i.e., within-participant manipulation at encoding only), and the fact that we had already observed consistent effects with similar sample sizes in the preceding experiments (i.e., 16–20 participants per *suppression* condition).

an acoustic component (i.e., linguistic suppression is both acoustic and rhythmic, while foot tapping is rhythmic only), which left open the possibility that any observed differences between linguistic suppression and our control secondary task could be due to auditory processing load. Similarly, we also judged it unsuitable to use control tasks involving use of the vocal cords and speech articulators, such as humming or clicking the tongue, which may have interfered with linguistic processing in inner speech. We, therefore, developed a novel control task of auditory monitoring that required participants to listen attentively for a pure tone and to press a foot response whenever they heard the tone. By matching the rhythmic and acoustic nature of linguistic suppression, limiting any additional motor activity to two brief, nonrepetitive occasions with an effector that was not required for reconstructing the models, and avoiding use of voicing and speech articulators, the auditory monitoring task met the criteria of a suitable control task.

In order to determine whether auditory monitoring matched linguistic suppression in terms of cognitive demands (i.e., whether they were of equivalent difficulty to perform), we pretested the tasks in a separate visual memory study using a sample of participants who did not take part in the main Experiment 3. The design and hypotheses were preregistered at <https://aspredicted.org/j9ye3.pdf>; its methods and results are summarized here but are fully reported, along with all materials, data, and code, at <https://osf.io/v8q47/>. We tested the auditory monitoring task using the visual patterns test (VPT; Della Sala et al., 1999; Miles et al., 1996; Wilson et al., 1987) of visual working memory. Specifically, we used a version of the VPT that was designed to minimize potential verbal coding (Brown et al., 2006), meaning that the stimuli do not have clear verbal labels and therefore cannot benefit from linguistic bootstrapping mechanisms. Since neither verbal labels nor auditory clicks and beeps were relevant to the visuospatial memory resources required for the VPT, any effects of linguistic suppression or auditory monitoring would simply be due to dual-task interference (i.e., the increased cognitive demands of performing a simultaneous secondary task). Critically, we expected both linguistic suppression and auditory monitoring tasks would have equivalent dual-task interference effects on the VPT, which would indicate that their cognitive demands were equivalent (i.e., that the tasks were of comparable difficulty; see Nedergaard et al., 2023).

An independent sample of participants ($N = 21$, determined via sequential hypothesis testing; 18 female; $M_{\text{age}} = 19.95$ years, $SD = 1.77$) was presented with patterns of black and white checkered squares of increasing complexity and then asked to reproduce these patterns from memory. There were three trials (patterns) per level of complexity (i.e., number of black squares in a grid), which increased incrementally from 2 to 15 squares. Each pattern was displayed on screen for 3 s during encoding, and participants were then presented with a blank grid for a maximum of 20 s and asked to reproduce (recall) the pattern by clicking on individual squares to change their color from white to black. Secondary task (three levels: no secondary task, linguistic suppression, auditory monitoring) was manipulated within participants. Participants performed linguistic suppression during encoding and recall exactly as in Experiment 2b. For the auditory monitoring task (see main experimental procedure for full details), participants heard the background clicks during encoding and recall, and responded via a foot response to a maximum of one tone per trial at either encoding or recall (the presence and timing of tones was varied pseudo-randomly

to ensure they were unpredictable—full details are available in the additional online materials). Participants were not required to respond to the clicks, only the tone. The experiment was conducted using PsychoPy3 (Version 2020.1.2; Peirce et al., 2019).

Working memory span (mean size of the last three correctly recalled patterns; minimum 3, maximum 15; Brown et al., 2006) was analyzed using hierarchical linear mixed-effects models and BFs for model comparison. When compared to baseline performance (i.e., no secondary task), carrying out a secondary task (i.e., collapsing across linguistic suppression and auditory monitoring) produced a small impairment of memory span (unstandardized $B = 0.53$, $SE = 0.24$, $t = 2.24$, $p = .025$) with equivocal evidence ($BF_{10} = 1.35$). That is, people remembered slightly fewer visual patterns due to the demands of carrying out a simultaneous secondary task (i.e., dual-task interference). Critically, when the two secondary tasks were compared, there was evidence against any difference between linguistic suppression and auditory monitoring ($BF_{10} = 0.18$; unstandardized $B = 0.24$, $SE = 0.27$, $t = 0.88$, $p = .381$), indicating that both tasks impaired visual memory to the same extent. Both linguistic suppression and auditory monitoring exerted comparable dual-task interference on visual memory, meaning they had equivalent demands as a secondary task when inner language cannot be used to support encoding of information. As such, the results of this pretest study confirmed that auditory monitoring was a suitable control task to compare against linguistic suppression in the present Experiment 3.

Design and Procedure

We used a within-participant design manipulating sequence length (models requiring seven, nine, and 11 steps to complete) and secondary task (linguistic suppression, auditory monitoring, and no secondary task). Each participant completed all three models in order of increasing sequence length, and secondary task was counterbalanced across sequence length using a Latin square design; that is, each participant experienced every secondary task condition but the order in which they experienced them varied. Secondary tasks in the present study were only performed during encoding to retest the consistent encoding effects observed in Experiments 1 and 2b.

For the auditory monitoring condition, participants heard the background “click” every 500 ms and two pure tones while they were watching the relevant instructional video. The timing of the tones varied pseudo-randomly between participants as they were allocated one of the nine different soundtracks for the instructional video of each model. Sound was played to participants through loudspeakers placed either side of the monitor and set at the same volume for each participant. Participants responded to the tones by pressing a foot response device containing a plastic squeaker that was attached to the underside of their dominant foot with an adjustable elastic strap. The experimenter instructed participants to listen carefully and respond every time they heard a tone, but participants did not know how many, or when, tones would be played. While participant watched the instructional videos, the experimenter listened and recorded whether the participant responded to both tones per video.

The experimenter explained each secondary task to participants immediately before they watched the relevant video, so participants did not know that they were going to be performing different secondary tasks before starting the experiment. Participants practiced

linguistic suppression before watching the relevant video with the experimenter, using a metronome as per Experiment 2b. Before performing auditory monitoring, the foot response device was attached to the participant's foot and the experimenter checked that the participant could use it correctly. The foot response device was then removed once the video had ended. The experimental procedure and setup were otherwise identical to Experiment 2b.

Data Preparation and Analysis

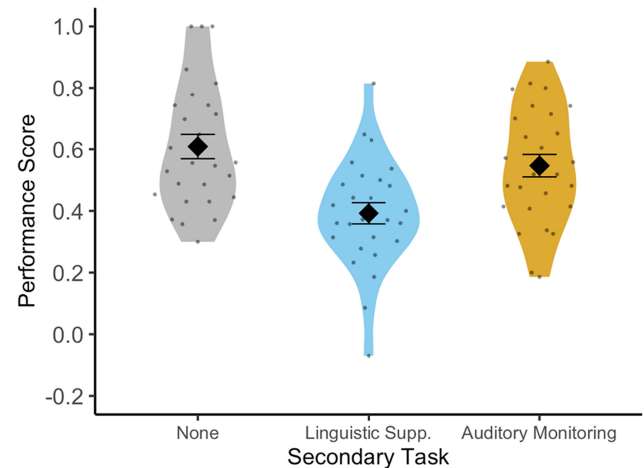
Video Coding and Scoring. Participant videos were coded in ELAN software and scored according to the final coding scheme for Experiment 2b, following exactly the same procedure. The second coder from Experiments 2a and 2b (still blind to the aims and conditions of the study) independently coded and scored a sample of 18 videos (22% of all videos, selected pseudo-randomly to cover a systematic cross-section of conditions and summed scores), to ensure that the scoring criteria could be objectively and consistently applied. The second coder independently scored the sample videos which were then compared to the first coder's using the same method as earlier Experiments. The coders reached sufficient agreement on the sample on the first attempt (completion: $k = 0.93$; errors: $k = 0.89$; multiple attempts: $k = 0.84$; serial order: $k = 0.81$), and the full data were analyzed without further discussion.

Statistical Analysis. All participants responded to both tones during the auditory monitoring task and so none were excluded from the analysis on this basis. One trial for the nine-step model was excluded from the completion time analysis as it was $>3 SD$ from the mean for that model. Participant performance score (measured in the same way as in Studies 2a and 2b) was analyzed using hierarchical linear mixed-effects models using the lme4 package (Version 1.1-23; Bates et al., 2015) in RStudio (Version 1.3.959; R Core Team, 2020). Sequence length comprised the number of events in a model and was analyzed as a continuous variable. The three levels of the secondary task (none, linguistic suppression, and auditory monitoring) were dummy-coded as follows with linguistic suppression as the overall reference level: secondary task (none = 1, auditory monitoring = 0, linguistic suppression = 0) and auditory monitoring (none = 0, auditory monitoring = 1, linguistic suppression = 0). Step 0 comprised the baseline model of participant as a random effect; Step 1 entered sequence length as a fixed effect; Step 2 added secondary task as a fixed effect; and Step 3 added auditory monitoring as a fixed effect. Successive hierarchical steps were compared using BFs, calculated based on the BIC (Wagenmakers, 2007). The Step 2 comparison against Step 1 tested whether memory performance was better or worse when a secondary task was performed, and the Step 3 comparison tested whether memory performance was specifically better or worse in the auditory monitoring condition compared to linguistic suppression. We also report marginal R^2 per step (calculated using the MuMIn package Version 1.43.17; Barton, 2017), coefficient statistics for the best-fitting model, and estimated marginal means per condition based on the final (i.e., most complex) model.

Results and Discussion

All reported analyses are confirmatory. Performance in constructing both models was overall good but again highly variable (see Figure 7). Performance scores ranged from 0.28 to 1.00 ($M = 0.62$,

Figure 7
Performance Score per Secondary Task Condition Across All Three Sequence Lengths in Experiment 3



Note. Diamonds represent means per condition with error bars of $\pm 1 SE$. Linguistic Supp. = linguistic suppression; OSF = Open Science Framework. See the online article for the color version of this figure.

$SD = 0.21$) for the seven-step fish model; from 0.09 to 0.89 ($M = 0.50$, $SD = 0.19$) for the nine-step bird model, and from -0.07 to 0.86 ($M = 0.43$, $SD = 0.19$) for the 11-step alien model.⁸

Performance scores were affected by sequence length, the presence of a secondary task, and the type of secondary task (see Table 5 for model comparisons). Strong evidence favored the Step 1 model over the Step 0 model, replicating previous findings from Experiments 2a to b that memory performance was affected by the number of events in a sequence. There was also strong evidence for the Step 2 model, showing that performing secondary tasks impaired memory performance and explained 10% of the variance in performance scores. Critically, evidence also strongly favored Step 3 over Step 2, indicating that model fit improved by an additional 9.2% of explained variance when the two secondary tasks were distinguished (i.e., auditory monitoring and linguistic suppression differentially affected performance score). Coefficients for the Step 3 model showed that, as predicted, memory performance was poorer for models with longer sequence lengths (unstandardized $B = -0.05$, $SE = 0.01$; standardized $\beta = -0.22$, $SE = 0.04$, $t = -5.27$, $p < .001$), and better when participants performed no secondary task compared to linguistic suppression (unstandardized $B = 0.22$, $SE = 0.04$; standardized $\beta = 1.03$, $SE = 0.17$, $t = 6.05$, $p < .001$). Critically, performance was also better in the auditory monitoring task compared to linguistic suppression (unstandardized $B = 0.15$, $SE = 0.04$; standardized $\beta = .74$, $SE = 0.17$, $t = 4.32$, $p < .001$). Marginal means (see Figure 7) indicated that, as predicted, performance was worst when linguistic suppression was performed at encoding, and best when no secondary task was performed.

That is, consistent with our predictions and replicating previous findings from Experiments 1 and 2b, memory for events was impaired

⁸ One participant achieved a negative performance score in the 11-step model due to skipping more than half the steps without attempting them.

Table 5

Overall Fit, Change in Fit, and Model Comparisons in Hierarchical Linear Regressions of Sequence Length and Secondary Task on Memory Performance Score in Experiment 3

Step	Parameter(s) added	R^2	ΔR^2	BF
0	Participant as random effect	0	—	—
1	Sequence length	.138	.138	130.52
2	Secondary task	.238	.100	83.77
3	Auditory monitoring	.329	.092	330.35

Note. Linguistic suppression is the reference level in coding of secondary task: the secondary task parameter at Step 2 therefore distinguishes between performing any secondary task versus none, while the auditory monitoring parameter at Step 3 distinguishes between it and linguistic suppression. BF = Bayes factor.

by linguistic suppression. This time, however, results confirmed that the effect of linguistic suppression on memory performance was not merely dual-task interference: linguistic suppression impaired performance more than a control secondary task (auditory monitoring) with equivalent demands. Rather, people's ability to remember and reproduce events was worse specifically due to limited access to inner language during event encoding.

General Discussion

We conducted three experiments to investigate the role of inner language in memory for events, testing the linguistic bootstrapping hypothesis that we habitually use linguistic labels as placeholders for complex sensorimotor representations of events. By employing linguistic suppression to disrupt access to inner language, we examined whether it affected people's ability to remember and recall nonverbal sequences of events. In Experiment 1, linguistic suppression at encoding and recall resulted in poorer memory performance, but completion time was not affected; however, the paradigm left open the possibility that the effect at recall was due to inner language supporting action planning rather than event memory. Following several improvements to the memory task and paradigm in Experiment 2a that allowed us to estimate maximum event memory capacity as approximately 7–8 events, Experiment 2b then replicated the effect of linguistic suppression at encoding in two different sequence lengths (within and exceeding memory capacity): disrupting access to inner language led to poorer performance, regardless of sequence length. However, this time memory performance was not affected when language was disrupted at recall, suggesting that inner language is critical to remembering events only during event encoding. Finally, Experiment 3 found linguistic suppression had a greater effect on memory for events than an auditory monitoring control task, which confirmed that—critically—the consistently observed effects of linguistic suppression at encoding were due to limited access to inner language and not merely the extra demands of a secondary task. Our findings therefore show that inner language enhances event encoding and memory even in nonverbal tasks where language is not explicitly involved (or even strictly necessary).

Overall, the results support the main predictions of the linguistic bootstrapping hypothesis. Most importantly, we observed consistent evidence that inner language supports and enhances the encoding of events, improving memory performance by increasing the number of events that can be remembered. When inner language was disrupted

during encoding, participants recalled 1–2 fewer events fully and accurately than when inner language was available. This effect translates to a significant real-world advantage for inner language in encoding complex perceptual and motor information. As people watch an event unfold, they automatically divide the continuous stream of experience into discrete segments—meaningful “chunks” of perceptual information that are separated at natural boundaries—and each segment is represented in memory as its own event model (Radvansky & Zacks, 2014; Zacks, 2020). What the present findings show is that event models are not merely perceptual/sensorimotor in nature but are also linguistic. By allowing aspects of an event to be represented as lightweight labels in place of rich and complex sensorimotor information, linguistic bootstrapping allows more information to be accurately represented within each event model while using fewer cognitive resources (and thus increasing memory capacity). That is, inner language enhances event memory by making the representation of each segmented event more efficient. It is for this reason that we find the critical effects in Experiments 1–3: people can recall a larger number of events, and reconstruct them more accurately, when inner language is available during event encoding compared to when it is not. These conclusions are consistent with the view that the representational format of event models integrates multiple sources of information including language (Zacks, 2020), and extends this role of language to incorporate inner language during nonverbal event cognition (i.e., without requiring overt verbalization; cf. Papafragou et al., 2008) and linguistic bootstrapping of representations within event models (i.e., enhancing representational efficiency). More broadly, the conclusions are also consistent with theories of concepts and cognition that argue language is an intrinsic part of the human conceptual system and therefore modulates how humans perceive and think about the world (Connell & Lynott, 2014; Louwse, 2011; Lupyan, 2012), and extend the evidence base for these theories to the domain of event memory.

By examining several measures of event memory performance, the present studies suggest that inner language benefits event memory in multiple ways. We primarily used a composite performance score that enabled us to assess participants' ability to remember an entire event sequence, by considering whether individual events were reproduced completely, accurately, without using trial and error, and in the correct sequential order. In exploratory analyses, we also examined a more traditional memory measure of the number of fully accurate and completed model steps, as well as the number of steps recalled completely but with errors (e.g., where a participant constructed the alien's head but inaccurately). Suppressing language at encoding saw a drop in all three measures, suggesting that inner language provides multiple levels of support to event memory. It can improve people's ability to perfectly recall a limited number of events (i.e., those recalled completely and accurately), but it also allows more events in general to be recalled regardless of accuracy (i.e., a limited number of events are recalled completely but with errors) or completeness (i.e., some events are recalled only partially, which is still better than nothing). Inner language therefore provides a mechanism to improve the quality and accuracy of event memory, but it can also act as a “good enough” mechanism to boost memory recall overall, even if some details are lost. Debrief questioning of participants after the experiment suggested that a rich variety of inner language may be used to support event representations. For example, participants reported mentally using a wide range of labels for objects and actions in the events, most commonly to describe the

features of the models and their construction such as colors, shapes, positions, and direction (e.g., red triangle, left/right, clockwise), as well as to describe the different segments of the models which comprised the subevents in the task (e.g., legs, feet, tail) and the overall model itself (e.g., fish). Indeed, they also reported using the names of existing, similar concepts to label pieces, segments, or patterns of the models (e.g., boat for curved Lego pieces; sweets (candy), desserts, or sports colors and national flags to remember combinations of colors). How event memory performance might be affected by individual participant strategies of inner language (e.g., different labels may affect event memory in different ways) would be an interesting area for future research.

Nevertheless, we did not find that inner language was consistently used during the recall process itself. Although we originally hypothesized that linguistic suppression at recall would impair memory performance because participants could not properly access linguistic labels in their memory of events, we realized after running Experiment 1 that an alternative explanation was possible. Experiment 1's birdhouse model had relatively few degrees of freedom: all the pieces available to participants were included in the final model, and the pieces themselves had limited affordances (i.e., slots and tabs) that constrained the possible ways they could be assembled. Hence, building the model "on the fly" (rather than by recalling from memory) was feasible simply by guessing which tab fit in which slot, which became progressively easier as each new piece was added. The kind of planning and problem-solving involved in this on-the-fly building strategy has previously been shown to be supported by inner language (e.g., the classic Tower of London task: Lidstone et al., 2010; Wallace et al., 2017). The effect of linguistic suppression we observed at recall in Experiment 1 could therefore have been due to inner language either affecting retrieval from event memory (as hypothesized) or affecting planning during on-the-fly building (the alternative possibility).

Thus, we abandoned the birdhouse model in subsequent experiments in favor of more complex models that minimized the chance of such on-the-fly strategies. The Lego models of Experiments 2–3 had much higher degrees of freedom: the large array of distractor bricks meant only a minority of available pieces were included in the final model, and the Lego bricks could fit together in many different ways. Hence, it would be extremely difficult for participants to build any of the models on the fly, and so we hypothesized that any effect of linguistic suppression at recall in Experiment 2b would have to be due to inner language affecting retrieval from event memory. However, no such recall effect emerged in Experiment 2b, which suggested that the recall effect in Experiment 1 was due to on-the-fly planning. To be clear, we are not claiming that participants performed Experiment 1 task solely through planning and building on the fly. The fact that encoding effects emerged suggested that, as we hypothesized, participants were using their event representation in memory (supported by inner language) to perform the task. Rather—and we acknowledge the explanation is speculative—we believe it is plausible that Experiment 1 participants could additionally use an on-the-fly building strategy during the recall stage, particularly in the later stages of construction. Future work could test this possibility by varying the number of distractor pieces available to participants while building a birdhouse-like model (i.e., manipulating degrees of freedom), and examining if the effect of linguistic suppression at recall reduced as the number of distractors increased.

The present pattern of effects echo similar findings in object memory (Dymarska et al., 2022), which also found a robust effect of

linguistic suppression at encoding but evidence against its effect at recall and interaction with encoding conditions. Considered together, the effects on memory for both objects and events suggest that inner language appears to support the formation of representations at the point of encoding, rather than retrieving them at the point of recall. That is, by a mechanism of linguistic bootstrapping, inner language enhances event memory by encoding the representation of each segmented event more efficiently. Indeed, the effects of inner language on event cognition may even go beyond linguistic bootstrapping. Although not tested in the present experiments, inner language may support event segmentation by making boundaries in an unfolding event more salient. Where a word label exists for a particular aspect of experience, its activation via inner language could plausibly help to identify and segregate this "chunk" of information within a stream of continuous experience, similar to how presenting a word label can improve visual perception of objects (e.g., Lupyan & Ward, 2013). Future research should examine whether and how inner language assists in the event segmentation process itself by boosting the identification of features that mark event boundaries.

Not all our original predictions were borne out. Linguistic suppression had no effect on the time taken to construct the models in any experiment, although we had expected the task to be more difficult and therefore slower when inner language was unavailable. This null effect may, in part, be due to individual differences in participant strategy: we observed quite large differences in the speed with which participants constructed the models (e.g., some participants spent time carefully checking different stages of their models while others did not), and this was not necessarily related to their memory performance. While performance score and completion time were inversely related in Experiments 1 and 2b (i.e., approximately 34%–38% shared variance), most variance in completion times was unrelated to memory performance. As there was no time limit for completing the models, participants were free to complete the task at their preferred speed. Consequently, completion time in open tasks involving complex event reproduction may not be a reliable indicator of memory performance. Alternatively, it may be the case that memory of 7–8 events supported by inner language can be recalled at much the same speed as the smaller number of events that can be held in memory when inner language is not available. Future research could investigate which possibility is more likely.

We initially predicted that linguistic labels would be more beneficial for longer sequences of events when memory capacity was strained, than for shorter sequences that fit within memory capacity. However, this prediction was not borne out in the present study: no interaction appeared between memory capacity and the role of inner language, and the effects of linguistic suppression were equivalent for sequences within and exceeding memory capacity. That is, inner language appeared to benefit memory performance regardless of sequence length. It is possible that inner language is employed in linguistic bootstrapping as soon as any event is represented, even when part of a short sequence that fits easily within memory capacity, hence leading inner language to enhance memory performance regardless of whether or not capacity is strained. An alternative possibility is that the events being encoded in our task were already sufficiently complex and demanding of cognitive resources to make inner language beneficial. That is, even the short event sequences had a mass of sensorimotor detail that could potentially be represented and could therefore benefit from linguistic bootstrapping, which may not have been the case if much simpler, sparser events

had been used. Future work could differentiate between these possible explanations by examining very simple events in varying sequence lengths to find out whether there exists a “tipping point” of inner language use in event memory. Nevertheless, the present results indicate that linguistic bootstrapping is generally a useful mechanism for inner language to support memory for events in both short and long sequences. Such a mechanism fits with Baddeley’s (2000) proposal that working memory has an episodic buffer that supports serial recall of integrated, cross-modal representations that are supported by information from long-term memory. This episodic buffer could therefore be the limited-capacity store that holds both sensorimotor representations of action sequences and the word labels activated via linguistic bootstrapping (see also Dymarska et al., 2022).

Finally, the present study provides an estimate of event memory capacity in a naturalistic event reproduction task. When inner language is available, Experiment 2a estimated memory capacity to be between seven and eight events—that is, memory performance was significantly poorer for sequences of events longer than seven or eight steps. It is important to note that this capacity estimate represents the maximum number of events that can be represented; it does not necessarily mean that seven to eight individual events will be recalled with perfect accuracy. Indeed, we found in Experiment 2b that people recalled only three events completely and accurately, or four events if we allowed for some degree of error. In other words, from a baseline memory capacity of seven to eight events when inner language is available to support event representations, it appears that people can remember approximately three events perfectly, one more complete event with some loss of accuracy, and a further three to four events with enough partial detail that they can be recalled to some extent. When access to inner language is disrupted during event encoding, however, these capacity estimates drop by one to two events. To the best of our knowledge, these estimates are the first for event memory capacity, and are broadly consistent with evidence that working memory capacity is generally found to be between three and seven “chunks” of information (e.g., Cowan, 2010; Simmering & Perone, 2013) or more when supported by information from long-term memory (e.g., up to 12 objects: Dymarska et al., 2022).

Conclusions

Our study provides consistent evidence that access to inner language enhances the encoding of event representations in memory, improving their overall accuracy and the number of events that can be recalled even when the events themselves are nonverbal. We provide the first estimate of event memory capacity, finding that people can accurately reproduce a maximum of seven to eight naturalistic events. The theoretical implications of these findings are that, firstly, they support the linguistic bootstrapping hypothesis, suggesting that linguistic labels are used as an efficient mechanism to form event representations in memory, acting as placeholders for complex bundles of connected sensorimotor information. Secondly, they add to existing theories of event memory, suggesting that inner language plays an important role in encoding events, helping to form an accurate and efficient working event model even when language is not being used in the task. Finally, they add to a growing body of literature demonstrating that inner language is an important cognitive tool that helps us to perform complex cognitive tasks.

References

- Alderson-Day, B., & Fernyhough, C. (2015). Inner speech: Development, cognitive functions, phenomenology, and neurobiology. *Psychological Bulletin*, *141*(5), 931–965. <https://doi.org/10.1037/bul0000021>
- Athanasopoulos, P., & Bylund, E. (2013). Does grammatical aspect affect motion event cognition? A cross-linguistic comparison of English and Swedish speakers. *Cognitive Science*, *37*(2), 286–309. <https://doi.org/10.1111/cogs.12006>
- Audacity Team. (2019). *Audacity®: Free audio editor and recorder* (Version 3.0.4) [Computer software]. <https://audacityteam.org/>
- Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. [https://doi.org/10.1016/S1364-6613\(00\)01538-2](https://doi.org/10.1016/S1364-6613(00)01538-2)
- Baddeley, A. (1966). The influence of acoustic and semantic similarity on long-term memory for word sequences. *Quarterly Journal of Experimental Psychology*, *18*(4), 302–309. <https://doi.org/10.1080/14640746608400047>
- Baddeley, A., Chincotta, D., & Adlam, A. (2001). Working memory and the control of action: Evidence from task switching. *Journal of Experimental Psychology: General*, *130*(4), 641–657. <https://doi.org/10.1037/0096-3445.130.4.641>
- Baddeley, A., & Hitch, G. J. (1974). Working memory. In G. Bower (Ed.), *Recent advances in learning and motivation* (Vol. 8, pp. 47–89). Academic Press.
- Bailey, H. R., Zacks, J. M., Hambrick, D. Z., Zacks, R. T., Head, D., Kurby, C. A., & Sargent, J. Q. (2013). Medial temporal lobe volume predicts elders’ everyday memory. *Psychological Science*, *24*(7), 1113–1122. <https://doi.org/10.1177/0956797612466676>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. <https://doi.org/10.1016/j.jml.2012.11.001>
- Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. In M. de Vega, A. Glenberg, & A. Graesser (Eds.), *Symbols and embodiment: Debates on meaning and cognition* (pp. 245–284). Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199217274.003.0013>
- Barton, K. (2017). *Package ‘MuMIn’*. <https://CRAN.R-project.org/package=MuMIn>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Bek, J., Blades, M., Siegal, M., & Varley, R. (2009). Linguistic processes in visuospatial representation: Clarifying verbal interference effects. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (pp. 2723–2728). Cognitive Science Society.
- Borghini, A. M., Barca, L., Binkofski, F., Castelfranchi, C., Pezzulo, G., & Tummolini, L. (2019). Words as social tools: Language, sociality and inner grounding in abstract concepts. *Physics of Life Reviews*, *29*, 120–153. <https://doi.org/10.1016/j.plrev.2018.12.001>
- Brandimonte, M. A., Hitch, G. J., & Bishop, D. V. M. (1992a). Influence of short-term memory codes on visual image processing: Evidence from image transformation tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(1), 157–165. <https://doi.org/10.1037/0278-7393.18.1.157>
- Brandimonte, M. A., Hitch, G. J., & Bishop, D. V. M. (1992b). Verbal recoding of visual stimuli impairs mental image transformations. *Memory & Cognition*, *20*(4), 449–455. <https://doi.org/10.3758/BF03210929>
- Brown, L. A., Forbes, D., & McConnell, J. (2006). Short article: Limiting the use of verbal coding in the visual patterns test. *Quarterly Journal of Experimental Psychology*, *59*(7), 1169–1176. <https://doi.org/10.1080/17470210600665954>
- Connell, L. (2019). What have labels ever done for us? The linguistic shortcut in conceptual processing. *Language, Cognition and Neuroscience*, *34*(10), 1308–1318. <https://doi.org/10.1080/23273798.2018.1471512>

- Connell, L., & Banks, B. (2024, February 08). *Inner language enhances memory for events*. <https://osf.io/v8q47/>
- Connell, L., & Lynott, D. (2014). Principles of representation: Why you can't represent the same concept twice. *Topics in Cognitive Science*, 6(3), 390–406. <https://doi.org/10.1111/tops.12097>
- Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science*, 19(1), 51–57. <https://doi.org/10.1177/0963721409359277>
- Cragg, L., & Nation, K. (2010). Language and the development of cognitive control. *Topics in Cognitive Science*, 2(4), 631–642. <https://doi.org/10.1111/j.1756-8765.2009.01080.x>
- Della Sala, S., Gray, C., Baddeley, A., Allamano, N., & Wilson, L. (1999). Pattern span: A tool for unwelding visuo-spatial memory. *Neuropsychologia*, 37(10), 1189–1199. [https://doi.org/10.1016/S0028-3932\(98\)00159-6](https://doi.org/10.1016/S0028-3932(98)00159-6)
- Dove, G. (2020). More than a scaffold: Language is a neuroenhancement. *Cognitive Neuropsychology*, 37(5–6), 288–311. <https://doi.org/10.1080/02643294.2019.1637338>
- Dunbar, K., & Sussman, D. (1995). Toward a cognitive account of frontal lobe function: Simulating frontal lobe deficits in normal subjects. *Annals of the New York Academy of Sciences*, 769(1), 289–304. <https://doi.org/10.1111/j.1749-6632.1995.tb38146.x>
- Dymarska, A., Connell, L., & Banks, B. (2022). Linguistic bootstrapping allows more real-world object concepts to be held in mind. *Collabra: Psychology*, 8(1), Article 40171. <https://doi.org/10.1525/collabra.40171>
- ELAN. (2019). *Nijmegen: Max Planck Institute for Psycholinguistics* (Version 5.8) [Computer software]. <https://tla.mpi.nl/tools/tla-tools/elan/>
- Emerson, M. J., & Miyake, A. (2003). The role of inner speech in task switching: A dual-task investigation. *Journal of Memory and Language*, 48(1), 148–168. [https://doi.org/10.1016/S0749-596X\(02\)00511-9](https://doi.org/10.1016/S0749-596X(02)00511-9)
- Fini, C., Zannino, G. D., Orsoni, M., Carlesimo, G. A., Benassi, M., & Borghi, A. M. (2022). Articulatory suppression delays processing of abstract words: The role of inner speech. *Quarterly Journal of Experimental Psychology*, 75(7), 1343–1354. <https://doi.org/10.1177/17470218211053623>
- Flecken, M., Athanasopoulos, P., Kuipers, J. R., & Thierry, G. (2015). On the road to somewhere: Brain potentials reflect language effects on motion event perception. *Cognition*, 141, 41–51. <https://doi.org/10.1016/j.cognition.2015.04.006>
- Flores, S., Bailey, H. R., Eisenberg, M. L., & Zacks, J. M. (2017). Event segmentation improves event memory up to one month later. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43(8), 1183–1202. <https://doi.org/10.1037/xlm0000367>
- Frank, M. C., Fedorenko, E., Lai, P., Saxe, R., & Gibson, E. (2012). Verbal interference suppresses exact numerical representation. *Cognitive Psychology*, 64(1–2), 74–92. <https://doi.org/10.1016/j.cogpsych.2011.10.004>
- Gimenes, G., Pennequin, V., & Mercer, T. (2016). What is the best strategy for retaining gestures in working memory? *Memory*, 24(6), 757–765. <https://doi.org/10.1080/09658211.2015.1049544>
- GraphPad Software. (n.d.). *Quantify agreement with kappa*. <https://www.graphpad.com/quickcalcs/kappa1/>
- He, H., Li, J., Xiao, Q., Jiang, S., Yang, Y., & Zhi, S. (2019). Language and color perception: Evidence from Mongolian and Chinese speakers. *Frontiers in Psychology*, 10, Article 551. <https://doi.org/10.3389/fpsyg.2019.00551>
- Hitch, G. J., Brandimonte, M. A., & Walker, P. (1995). Two types of representation in visual memory: Evidence from the effects of stimulus contrast on image combination. *Memory & Cognition*, 23(2), 147–154. <https://doi.org/10.3758/BF03197217>
- Imbo, I., & LeFevre, J.-A. (2010). The role of phonological and visual working memory in complex arithmetic for Chinese- and Canadian-educated adults. *Memory & Cognition*, 38(2), 176–185. <https://doi.org/10.3758/MC.38.2.176>
- Jaroslawska, A. J., Gathercole, S. E., & Holmes, J. (2018). Following instructions in a dual-task paradigm: Evidence for a temporary motor store in working memory. *Quarterly Journal of Experimental Psychology*, 71(11), 2439–2449. <https://doi.org/10.1177/1747021817743492>
- Kurby, C. A., & Zacks, J. M. (2011). Age differences in the perception of hierarchical structure in events. *Memory & Cognition*, 39(1), 75–91. <https://doi.org/10.3758/s13421-010-0027-2>
- Larsen, J. D., & Baddeley, A. (2003). Disruption of verbal STM by irrelevant speech, articulatory suppression, and manual tapping: Do they have a common source? *The Quarterly Journal of Experimental Psychology Section A*, 56(8), 1249–1268. <https://doi.org/10.1080/02724980244000765>
- Lidstone, J. S., Meins, E., & Fernyhough, C. (2010). The roles of private speech and inner speech in planning during middle childhood: Evidence from a dual task paradigm. *Journal of Experimental Child Psychology*, 107(4), 438–451. <https://doi.org/10.1016/j.jecp.2010.06.002>
- Logie, R. H., & Baddeley, A. D. (1987). Cognitive processes in counting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(2), 310–326. <https://doi.org/10.1037/0278-7393.13.2.310>
- Logie, R. H., Gilhooly, K. J., & Wynn, V. (1994). Counting on working memory in arithmetic problem solving. *Memory & Cognition*, 22(4), 395–410. <https://doi.org/10.3758/BF03200866>
- Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, 3(2), 273–302. <https://doi.org/10.1111/j.1756-8765.2010.01106.x>
- Lupyan, G. (2009). Extracommunicative functions of language: Verbal interference causes selective categorization impairments. *Psychonomic Bulletin & Review*, 16(4), 711–718. <https://doi.org/10.3758/PBR.16.4.711>
- Lupyan, G. (2012). Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in Psychology*, 3, Article 54. <https://doi.org/10.3389/fpsyg.2012.00054>
- Lupyan, G., & Ward, E. J. (2013). Language can boost otherwise unseen objects into visual awareness. *Proceedings of the National Academy of Sciences*, 110(35), 14196–14201. <https://doi.org/10.1073/pnas.1303312110>
- Miles, C., Morgan, M. J., Milne, A. B., & Morris, E. D. M. (1996). Developmental and individual differences in visual memory span. *Current Psychology: A Journal for Diverse Perspectives on Diverse Psychological Issues*, 15(1), 53–67. <https://doi.org/10.1007/BF02686934>
- Minda, J. P., Desroches, A. S., & Church, B. A. (2008). Learning rule-described and non-rule-described categories: A comparison of children and adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(6), 1518–1533. <https://doi.org/10.1037/a0013355>
- Mitsuhashi, S., Hirata, S., & Okuzumi, H. (2018). Role of inner speech on the Luria hand test. *Cogent Psychology*, 5(1), Article 1449485. <https://doi.org/10.1080/23311908.2018.1449485>
- Morin, A. (2018). The self-reflective functions of inner speech: Thirteen years later. In P. Langland-Hassan & A. Vicente (Eds.), *Inner speech: New voices* (pp. 276–298). Oxford University Press.
- MusicRadar. (n.d.). *SampleRadar: 301 free percussion loops*. <https://www.musicradar.com/news/sampleradar-percussion-loops-samples-1>
- Nakabayashi, K., & Burton, A. M. (2008). The role of verbal processing at different stages of recognition memory for faces. *European Journal of Cognitive Psychology*, 20(3), 478–496. <https://doi.org/10.1080/09541440.801946174>
- Nedergaard, J., Wallentin, M., & Lupyan, G. (2023). Verbal interference paradigms: A systematic review investigating the role of language in cognition. *Psychonomic Bulletin & Review*, 30(2), 464–488. <https://doi.org/10.3758/s13423-022-02144-7>
- Newton, D. (1976). Foundations of attribution: The perception of ongoing behavior. In J. H. Harvey, W. J. Ickes, & K. F. Kidd (Eds.), *New directions in attribution research* (pp. 223–248). Lawrence Erlbaum Associates.
- Papafragou, A., Hulbert, J., & Trueswell, J. (2008). Does language guide event perception? Evidence from eye movements. *Cognition*, 108(1), 155–184. <https://doi.org/10.1016/j.cognition.2008.02.007>
- Papafragou, A., Massey, C., & Gleitman, L. (2002). Shake, rattle, 'n' roll: The representation of motion in language and cognition. *Cognition*, 84(2), 189–219. [https://doi.org/10.1016/S0010-0277\(02\)00046-X](https://doi.org/10.1016/S0010-0277(02)00046-X)
- Pearce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). Psychopy2: Experiments in

- behavior made easy. *Behavior Research Methods*, 51(1), 195–203. <https://doi.org/10.3758/s13428-018-01193-y>
- Pelizzon, L., Brandimonte, M. A., & Favretto, A. (1999). Imagery and recognition: Dissociable measures of memory? *European Journal of Cognitive Psychology*, 11(3), 429–443. <https://doi.org/10.1080/713752323>
- Phillips, L. H., Wynn, V., Gilhooly, K. J., Della Sala, S., & Logie, R. H. (1999). The role of memory in the Tower of London task. *Memory*, 7(2), 209–231. <https://doi.org/10.1080/741944066>
- Radvansky, G. A., & Zacks, J. M. (2014). *Event cognition*. Oxford University Press.
- R Core Team. (2020). *R: A language and environment for statistical computing* (Version 1.3.939) [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Roberson, D., & Davidoff, J. (2000). The categorical perception of colors and facial expressions: The effect of verbal interference. *Memory & Cognition*, 28(6), 977–986. <https://doi.org/10.3758/BF03209345>
- Robert, N. D., & LeFevre, J.-A. (2013). Ending up with less: The role of working memory in solving simple subtraction problems with positive and negative answers. *Research in Mathematics Education*, 15(2), 165–176. <https://doi.org/10.1080/14794802.2013.797748>
- Rubin, D. C., & Umanath, S. (2015). Event memory: A theory of memory for laboratory, autobiographical, and fictional events. *Psychological Review*, 122(1), 1–23. <https://doi.org/10.1037/a0037907>
- Saeki, E., & Saito, S. (2004). Effect of articulatory suppression on task-switching performance: Implications for models of working memory. *Memory*, 12(3), 257–271. <https://doi.org/10.1080/09658210244000649>
- Santin, M., Van Hout, A., & Flecken, M. (2021). Event endings in memory and language. *Language, Cognition and Neuroscience*, 36(5), 625–648. <https://doi.org/10.1080/23273798.2020.1868542>
- Sargent, J. Q., Zacks, J. M., Hambrick, D. Z., Zacks, R. T., Kurby, C. A., Bailey, H. R., Eisenberg, M. L., & Beck, T. M. (2013). Event segmentation ability uniquely predicts event memory. *Cognition*, 129(2), 241–255. <https://doi.org/10.1016/j.cognition.2013.07.002>
- Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, 22(2), 322–339. <https://doi.org/10.1037/met0000061>
- Seitz, K., & Schumann-Hengsteler, R. (2002). Phonological loop and central executive processes in mental addition and multiplication. *Psychologische Beiträge*, 44(2), 275–302.
- Shebani, Z., & Pulvermüller, F. (2013). Moving the hands and feet specifically impairs working memory for arm- and leg-related action words. *Cortex*, 49(1), 222–231. <https://doi.org/10.1016/j.cortex.2011.10.005>
- Simmering, V. R., & Perone, S. (2013). Working memory capacity as a dynamic process. *Frontiers in Psychology*, 3, Article 567. <https://doi.org/10.3389/fpsyg.2012.00567>
- Skordos, D., Bunger, A., Richards, C., Selimis, S., Trueswell, J., & Papafragou, A. (2020). Motion verbs and memory for motion events. *Cognitive Neuropsychology*, 37(5–6), 254–270. <https://doi.org/10.1080/02643294.2019.1685480>
- Smyth, M. M., & Pendleton, L. R. (1989). Working memory for movements. *The Quarterly Journal of Experimental Psychology Section A*, 41(2), 235–250. <https://doi.org/10.1080/14640748908402363>
- ter Bekke, M., Özyürek, A., & Ünal, E. (2022). Speaking but not gesturing predicts event memory: A cross-linguistic comparison. *Language and Cognition*, 14(3), 362–384. <https://doi.org/10.1017/langcog.2022.3>
- Trbovich, P. L., & LeFevre, J.-A. (2003). Phonological and visual working memory in mental addition. *Memory & Cognition*, 31(5), 738–745. <https://doi.org/10.3758/BF03196112>
- Trueswell, J. C., & Papafragou, A. (2010). Perceiving and remembering events cross-linguistically: Evidence from dual-task paradigms. *Journal of Memory and Language*, 63(1), 64–82. <https://doi.org/10.1016/j.jml.2010.02.006>
- van der Loo, M. (2014). The stringdist package for approximate string matching. *The R Journal*, 6(1), 111–122. <https://doi.org/10.32614/RJ-2014-011>
- Vygotsky, L. B. (1986). *Thought and language* (A. Kozulin, Trans.; Rev. ed.). MIT Press. (Original work published 1934)
- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, 14(5), 779–804. <https://doi.org/10.3758/BF03194105>
- Wallace, G. L., Peng, C. S., & Williams, D. (2017). Interfering with inner speech selectively disrupts problem solving and is linked with real-world executive functioning. *Journal of Speech, Language, and Hearing Research*, 60(12), 3456–3460. https://doi.org/10.1044/2017_JSLHR-S-16-0376
- Wilson, J. L., Scott, J. H., & Power, K. G. (1987). Developmental differences in the span of visual memory for pattern. *British Journal of Developmental Psychology*, 5(3), 249–255. <https://doi.org/10.1111/j.2044-835X.1987.tb01060.x>
- Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian Blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, 104(19), 7780–7785. <https://doi.org/10.1073/pnas.0701644104>
- Wingfield, C., & Connell, L. (2022). Understanding the role of linguistic distributional knowledge in cognition. *Language, Cognition and Neuroscience*, 37(10), 1220–1270. <https://doi.org/10.1080/23273798.2022.2069278>
- Zacks, J. M. (2020). Event perception and memory. *Annual Review of Psychology*, 71(1), 165–191. <https://doi.org/10.1146/annurev-psych-010419-051101>
- Zacks, J. M., Speer, N. K., Vettel, J. M., & Jacoby, L. L. (2006). Event understanding and memory in healthy aging and dementia of the Alzheimer type. *Psychology and Aging*, 21(3), 466–482. <https://doi.org/10.1037/0882-7974.21.3.466>
- Zacks, J. M., & Swallow, K. M. (2007). Event segmentation. *Current Directions in Psychological Science*, 16(2), 80–84. <https://doi.org/10.1111/j.1467-8721.2007.00480.x>
- Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin*, 127(1), 3–21. <https://doi.org/10.1037/0033-2909.127.1.3>
- Zacks, J. M., Tversky, B., & Iyer, G. (2001). Perceiving, remembering, and communicating structure in events. *Journal of Experimental Psychology: General*, 130(1), 29–58. <https://doi.org/10.1037/0096-3445.130.1.29>
- Zeithamova, D., & Maddox, W. T. (2007). The role of visuospatial and verbal working memory in perceptual category learning. *Memory & Cognition*, 35(6), 1380–1398. <https://doi.org/10.3758/BF03193609>

Received February 13, 2023

Revision received February 13, 2024

Accepted February 14, 2024 ■