The emergence of all-or-none retrieval of chunks in verbal serial recall

Amirhossein Shafaghat Ardebili[1], Yang S. Liu[2], and Jeremy B. Caplan[3,1]

[1]Neuroscience and Mental Health Institute, University of Alberta, Edmonton, AB, T6G 2E1, Canada

[2]Department of Psychiatry, University of Alberta, Edmonton, AB, T6G 2R3, Canada

[3]Department of Psychology, University of Alberta, Edmonton, AB, T6G 2E9, Canada

Abstract

People often subdivide a list into smaller pieces, called chunks. Some theories of serial recall assume memories are stored hierarchically, with all-or-none retrieval of chunks, but most mathematical models avoid hierarchical assumptions. Johnson (1969) found steep drops in errors following correct recalls (transitional-error probabilities) within putative chunks during multi-trial letter-list learning, and viewed this as evidence for all-or-none retrieval. Here we test whether all-or-none retrieval occurs in lists studied only once. In serial recall of six-word lists (Experiment 1), transitional-error probabilities were inconsistent with all-or-none retrieval, both when participants were instructed to subdivide and when temporal grouping induced subdivision. Curiously, the same analysis of previous temporally grouped nine-letter lists produced compelling evidence for all-or-none retrieval, which may result from recoding rather than the formation of chunks. In Experiment 2, participants were pre-trained on three-word chunks. For nine-word lists constructed from those trained chunks, transitional-error probabilities exhibited more pronounced evidence of all-or-none retrieval. Nearly all effects reversed with post-cued backward recall, suggesting mechanisms that play out over the course of recall rather than encoding of the list. In sum, subdivided lists do not result in hierarchical memories after a single study trial, although they may emerge in lists formed from chunks that are previously learned as such. This suggests a continuous transition from non-hierarchical subdivision of lists to all-or-none retrieval over the course of chunk formation.

Keywords: Serial recall, backward recall, temporal grouping, chunking, hierarchical memory

Introduction

Subdividing a list is thought to make the list easier to remember and has been repeatedly shown to improve serial recall accuracy. Proposed mechanisms include temporal

or positional distinctiveness, rehearsal patterns, information-compression and hierarchical organization of the list in memory, as well as acting on redintegration during recall (e.g., Farrell, 2012; Gobet et al., 2016; Henson, 1998; Johnson, 1969; Norris & Kalm, 2021). In fact, all factors might contribute to advantages due to subdividing lists. To elaborate: 1) Distinctiveness. Some positional/ordinal coding models, such as SIMPLE (Brown et al., 2007), the Start-End Model (Henson, 1998) and OSCAR (Brown et al., 2000) build memory of a serial list by associating each list-item to a separate representation of order. These models have produced advantages for subdivided lists by invoking a second level of positional coding, either position within chunks/groups or else position of the chunk/group, itself. 2) Rehearsal patterns. Subdivision may influence how participants rehearse items, and which items get rehearsed together (Ward & Tan, 2023; Wickelgren, 1967) but Ryan (1969b) came to a different conclusion. 3) Information compression. Grouping items can sometimes offer an advantage by reducing the information-content of the list, particularly when stimuli are not strictly random, but include regular patterns that are more frequent than others (e.g., Norris & Kalm, 2021). 4) Hierarchical organization of memory. Subdividing a list has been proposed to result in a memory that is stored in a hierarchical manner; the list is composed of sublists (chunks), and each chunk is, in turn, comprised of items (e.g., Farrell, 2012; Johnson, 1969; Lee & Estes, 1981). 5) Redintegration or effects during the course of recall. Subdividing a list may facilitate the recall process by providing ways to redintegrate (clean up a messy retrieval) rather than (solely) influencing the information that is encoded (e.g., Norris & Kalm, 2021). 6) Recoding. Some materials might be recoded, such as converting sequences of three binary elements (0 or 1) into octal numbers, as suggested by Miller (1956), although if stimuli are truly random, this evidently does not produce an advantage that offsets the extra overhead processing required for the conversion (Glanzer & Fleishman, 1967). For lists of lists of letters, as Miller (1956) noted, recoding might occur by finding the most similar word (such as vanity license plates; BLK may be remembered as "black") or by replacement with known acronyms (FBI, BLM). 7) Finally, participants have some flexibility to alter their order of report (Cowan et al., 2002; Grenfell-Essam & Ward, 2012; Kahana et al., 2010; Ward & Tan, 2019; Watkins & Bloom, 1999). It is possible that subdividing a list makes it easier for participants to strategically control retrieval order. Consistent with this, Spurgeon et al. (2015) found that participants often initiated recall from the start of the most recent group, when word lists were studied temporally grouped.

Our curiosity stems from early ideas about chunking. When he introduced the word "chunking," Miller (1956) initially thought memory can become in some sense hierarchical; the participant remembers the list by remembering the sequence of chunks and recalling the items within each chunk. This concept aligns with models that assume a list is stored

Jeremy B. Caplan https://orcid.org/0000-0002-8542-9900. Yang S. Liu https://orcid.org/0000-0003-0406-8056. Amirhossein Shafaghat Ardebili https://orcid.org/0000-0001-9574-0623.
Corresponding author: Jeremy B. Caplan. Department of Psychology, Biological Sciences Building, University of Alberta, Edmonton, Alberta T6G 2E9, Canada, E-mail: jcaplan@ualberta.ca, Tel: +1.780.492.5265, Fax: +1.780.492.1768.
Supported in part by the Natural Sciences and Research Council of Canada. The data can be retrieved from https://osf.io/hf2ka. Thanks to Madhawa Alahakoon and Briana Kroeker for work on earlier experiments that led to these ones.

and retrieved in a strictly hierarchical manner (Farrell, 2012), but also with recoding, if the participant must first access the code (the concept of FBI) and then unpack its associated elements (F, B and I).

Johnson (1969) proposed a way to test this idea (expanded in Johnson, 1970). If a chunk is retrieved all-or-none, then the first item might or might not be recalled, but if it is recalled, the remaining items of the chunks should be recalled with very low error rates. Johnson reasoned that an error spike pattern should be observed; error probability could be substantial leading into a chunk, but should level off at a very low rate for transitions within the chunk. Johnson suggested examining what he called transitional-error probability (TEP) plots, the proportion of errors following correct recalls as a function of output transition.

To understand why Johnson expected spikes, consider the following examples. Suppose a participant studied a list A B C D E F, where letters stand in for items. Trivially, if the participant recalls every item incorrectly, no data will contribute to the TEP calculation. Also trivially, if the participant recalls the list perfectly (A B C D E F), TEP (including the initiation error probability at the start of recall) will be 0 0 0 0 0 0. Thus, ceiling and floor performance do not produce any variation of TEP across output transition. Next consider a recall sequence A B C X Y Z. The TEP sequence will be 0 0 0 1 _ _ (where _ denotes a missing value, left out of the calculation. Compare with a recall sequence X Y C D E F. The values contributing to the TEP function are 1 _ _ 0 0 0. Now, the TEP is accumulated over many lists. Suppose the serial position curve (accuracy as a function of serial position) is a roughly descending function, 0.90 0.85 0.80 0.75 0.70 0.65. If recalls are independent of one another, the TEPs will simply be 1–accuracy, thus: 0.10 0.15 0.20 0.25 0.30 0.35, with no spikes between the two three-item subsequences. This is because independence means that accuracy on the current recall is unrelated to accuracy of the prior recall, so conditionalizing on the prior response being correct will yield the same error probability, on average, as if one were to conditionalize on the prior response being incorrect. If the two three-item subsequences are often retrieved all-or-none, a sequential dependency is introduced. Of most interest is the transition between the two putative chunks, from item 3 to item 4. If chunks, themselves, are retrieved independently of one another, the TEP from $3 \rightarrow 4$ will be 1–accuracy of the fourth item (0.25 again for this example). But given that the fourth item is correct, both item 5 and item 6 are likely to be correct, thus the final two TEPs will be 0 for such all-or-none retrievals. The spike that Johnson anticipated is because all-or-none retrieval substantially deviates from independence of successive retrievals, and with a tell-tale pattern: reduced error rate within chunks, assuming the chunk is initiated, but little influence on that chunk-initiation probability, itself.

Johnson (1969, 1970) found evidence of such TEP spikes. However, when we plotted the data, the evidence was not as pronounced as one might imagine (Figure 1). Peaks are indeed visible at chunk boundaries (3→4 for 3–4 patterned lists and 2→3 and 5→6 for the 2–3–2 patterned lists). However, the drop from the spike is underwhelming; within-chunk transitional errors are not less than half the error rate to initiate the chunk with the exception of the last two-item chunk in the 2–3–2 lists. This is hardly the dramatic characteristic that "all-or-none" would imply.

But Johnson's methods were particular. Participants mastered lists over multiple study/serial-recall cycles. Johnson only reported transitional-error probabilities *aggregated*

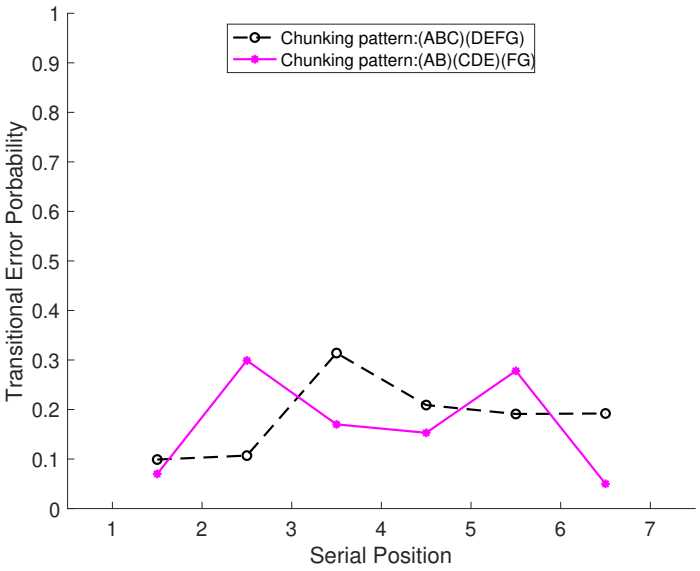

Figure 1

Transitional-Error Probabilities (probability of an error response whenever the prior response was correct) replotted from the table for the first experiment reported by Johnson (1970), for groups composed of groups of 3 and 4 letters each (black dashed plot) or composed of groups of 2, 3, and 2 letters each (purple solid plot). Note that the data points are for transitions, and are plotted between the corresponding serial positions.

over cycles, but on the first cycle, one would expect mostly errors, whereas on late cycles, accuracy should approach ceiling. The greatest sensitivity would presumably be in the middle cycles, so the subtlety of the spike pattern might be due to the early and late cycles washing out an underlying effect that is more clearly suggestive of all-or-none retrieval. Moreover, Johnson's lists were presented simultaneously and with spaces delineating chunk boundaries. This deviates from the bulk of serial recall studies which present items one at a time and without spatial cueing of chunk boundaries.

Meanwhile, as summarized above, there are other accounts of subdividing once-presented sequential lists tested with serial recall that have fit behavioural data. However, these accounts largely make no reference to hierarchical organization of memory (but see J. R. Anderson and Matessa, 1997; Farrell, 2012; Murdock, 1993, 1995 although Murdock, 2005 later rejected chunking) or all-or-none retrieval. Note that in the Perturbation Model (Lee & Estes, 1981), chunks modulated probabilities of perturbations an item could undergo, but did not assume chunks were retrieved all-or-none. Items were still retrieved independently, as was the case for J. R. Anderson and Matessa, 1997. Considering this previous research, we were left wondering if anything like all-or-none retrieval of chunks might occur in these more dominant paradigms, especially when the participant studies the target list just once.

Despite Johnson's lukewarm evidence of all-or-none retrieval, there are certainly times when memory is accessed hierarchically. Consider the following thought experiment: If a participant were given a list of words corresponding to song names, having previously

memorized the full lyrics to each: Hallelujah Fame Yesterday Respect. When asked to reproduce the lyrics of all those songs in that order, the participant would likely retrieve the entire list of tens of words in perfect, or near-perfect order— far exceeding word span. If we presented the following four-word list: Fame Respect Hallelujah Yesterday (a different "chunk" order), the participant could still probably recall the entire list of lyrics to the four songs in the correct order. Each "chunk" (song) would presumably be recalled nearly all-or-none, as Johnson envisioned— errors would be mostly omission or relocation of an entire "chunk" (with the occasional item-error too, as Johnson noted). Such lists would embody the "opaque label" property (Johnson, 1969). Arguably, words themselves could be viewed as chunks (Miller, 1956), and moreover, chunks that are retrieved all-or-none— the vast published data on serial recall of word lists stands as proof of principle. The four-song list examples could be viewed as lists retrieved via a 4-level hierarchy (list, song, word, letter).

The idea that all-or-none retrieval *never* occurs therefore seems untenable. Perhaps with substantial amount of repeated study of a subdivided list, memory becomes hierarchical and retrieval of each chunk is approximately all-or-none. Johnson's procedure might have revealed the gradual emergence of chunks. The question remains whether all-or-none retrieval of chunks can occur at the other end of the continuum, without overlearning of lists or chunks. We wondered whether, when participants subdivide a list, their memories become approximately hierarchical and chunks retrieved approximately all-or-none following only a single exposure to the list.

Besides multiple training cycles for each list, Johnson's lists were composed of letters. Letters (and digits) are conducive to recoding. Recoding is well known to be exploited by memory champions and highly trained memorizers, such as the case study by Ericsson et al. (1980) who recoded three-digit sequences as running times and two-digit sequences as ages (Ericsson et al., 2004). As already noted, letters can be recoded as acronyms or words (Miller, 1956). Lists of randomly selected nouns are not so easily recoded, although the so-called "story mnemonic," where the memorizer constructs a brief story to combine sequences of a few words, is effective (Bower & Clark, 1969; Worthen & Hunt, 2008) but requires training and practice. We wondered if evidence of all-or-none retrieval could be found when recoding was unlikely, with randomly constructed lists of words (rather than digits or letters).

Finally, we use backward serial recall as a tool to narrow down mechanisms of subdivision in our paradigms. In backward serial recall, the participant is asked to retrieve the list in reverse-presentation order. Theories of serial-order memory largely do not address backward recall. Only occasionally has backward recall been implemented in mathematical models (Bireta et al., 2010; Liu & Caplan, 2020). Some dissociations of experimental manipulations interact with recall direction. For example, Li and Lewandowsky (1993) reported findings suggesting backward recall is more dependent on visuospatial coding than forward recall (but see Guitard and Saint-Aubin, 2021; Guitard et al., 2019 for a more nuanced view). Meanwhile, many factors do not differ by recall direction, including temporal grouping in lists of consonants (Liu & Caplan, 2020), and under some conditions like reconstruction, where only order needs to be reconstructed but not items, near-equivalence can even be observed (Farrand & Jones, 1996). Particularly with visual presentation (Madigan, 1971), serial-position curves are almost mirror-images of one another; in other words, accuracy is primarily a function of output order (recall position) rather than study order

(serial position). This suggests that, in line with the model implementations of Bireta et al. (2010) and Liu and Caplan (2020), many of the underlying cognitive mechanisms may be the same. Comparing backward to forward recall direction can thus disambiguate whether a given experimental manipulation influences memory as a function of serial position, implying processes during study, or output position, implying processes that materialize over the course of the recall sequence. If an experimental factor interacts with both recall direction and output position, that would leave open the possibility that the factor influences memory during the study phase, as a function of serial position. But if the three-way interaction of the factor with recall direction and output position is not observed, that would point to processes developing over the course of recall, such as due to output interference or output encoding effects. To our knowledge, only Liu and Caplan (2020) have compared grouping directly between backward and forward recall directions (but see Anders and Lillyquist, 1971 who reported inter-response times for a single participant who reported grouping lists in twos) and found no three-way interaction involving output order. We asked if we would replicate this result with lists of words, implying mechanisms of list-subdivision acting primarily as a function of output, rather than serial position. Farrell (2012), in his hierarchical model, assumed that chunks would be retrieved preferably in the forward direction. For this reason, that model, as published, would predict an interaction with output-order. Insofar as Farrell's model is supported, if we extend to word lists the lack of interaction of grouping with recall direction and output position found by Liu and Caplan (2020) for consonant lists, that would call for a reconsideration of the forward-retrieval assumption.

As an aside, we analyze within-list errors, with a focus on interpositions, Henson's term for a transposition of an item to the wrong position, but nonetheless at the correct within-group position. These types of errors are expected to be frequent based on models that assume recall is carried out by cueing with some sort of positional information, and that includes within-group position when a list is grouped (e.g., Brown et al., 2007; Farrell, 2012; Henson, 1998). Liu and Caplan (2020) found high rates of interpositions in temporally grouped consonant lists, but only in the forward recall direction. This questioned the idea that serial recall of the consonant data were supported by positional cueing. By examining interposition rates in word lists, both temporally and subjectively grouped, we planned to test the validity of position- or order-based retrieval in the current experiments.

Experimental design and rationale. In two new experiments and a re-analysis of previous data, we ask whether all-or-none retrieval can be confirmed for sequentially presented lists that are studied just once. Both new experiments use the novel combination of word lists with a manipulation of recall direction to disambiguate effects of output order versus study order, and with a manipulation of temporal grouping (Experiment 1) or subjective chunking (both experiments). The first new experiment evaluates TEPs in lists of words, recalled in both the forward and backward direction and with respect to temporal grouping and instructions to chunk the list. We compare these findings with new TEP analyses of previously published consonant-lists tested with both forward and backward serial recall (Liu & Caplan, 2020) which, like Johnson's materials, might be conducive to recoding. Because the word-list results did not show convincing evidence of all-or-none retrieval, we conducted a second experiment inspired by Chen and Cowan (2005) and Thalmann et al. (2019), where chunks were pre-trained and report TEPs for recall of lists composed of those pre-trained chunks, finally suggesting that all-or-none retrieval emerges after a small

amount of prior study of chunks. We discuss the cumulative evidence about TEPs with respect to the range of models of list subdivision.

Experiment 1

In the first experiment, participants studied lists of sequentially presented words, with one serial recall attempt of each list. Lists were presented at a uniform rate (Control) or with pauses between chunks (Temporal Grouping) or uniformly but with explicit instructions to subdivide the list (Subjective Chunking). Recall was requested in either the forward or backward direction, manipulated between subjects. This experiment was closely modelled after Experiment 1 of Liu and Caplan (2020), preserving the same presentation rates in the control and temporally grouped conditions, respectively, including equating total study time across conditions. To compensate for the increased complexity of words compared to letters, we reduced the list length from nine to six. When subdivided, the list comprised two groups of three words each, thus group size was the same as in Liu and Caplan (2020). We added the Subjective Chunking condition in case temporal grouping was too subtle or failed to give participants the idea to conceptualize the list as subdivided. Inclusion of backward as well as forward recall directions allows us to test whether features of the data are primarily functions of serial position or of output position as was found by Liu and Caplan (2020). Interestingly, temporal grouping has been studied with lists of digits and letters but only twice, to our knowledge, lists of words (Kowialiewski et al., 2021; Spurgeon et al., 2015); but see Hitch et al. (1996) who did not report those serial-position effects. The temporal grouping/forward condition of Experiment 1 thus can also be viewed to some degree as a replication attempt of these, speaking to the robustness of temporal grouping as a manner of inducing subdivision of word lists.

Methods

The data from both experiments are available on osf.io and neither experiment was pre-registered.

Participants

A total of 308 participants were recruited from the Prolific website (www.prolific.co) and completed the experiment on the Pavlovia website (www.pavlovia.org) in exchange for GBP £7. The original sample size target 180 participants (30 participants per condition, modelled on Liu and Caplan, 2020 with similar design but anticipating greater sensitivity due to fewer serial positions); however, because counterbalancing across participants is not supported by Pavlovia, we assigned conditions at random, resulting in different recruited sample sizes per condition. We also ran additional participants in groups of 10 (plus additional participants in an effort to account for excluded participants throughout data-collection) in an effort to obtain a conclusive Bayes Factor for our main effect of interest, the three-way interaction accuracy effect ($BF_{\text{inclusion}}$ >3:1 or <1:3), the critical finding in Liu and Caplan (2020). A robustness check of the effect size as suggested by R. B. Anderson et al. (2022), plotted in Figure S11a, shows that the η_p^2 was fairly stable toward the end of data-collection. After excluding 37 participants with floor (fewer than 1 word recalled per list on average) or ceiling (more than 5 words recalled per list on average), final included

numbers were: Forward/Control: 42 (out of 48), Backward/Control: 43 (out of 49), Forward/Subjective chunking: 68 (out of 75); Backward/Subjective chunking: 31 (out of 39), Forward/Temporal grouping: 40 (out of 44), Backward/Temporal grouping: 47 (out of 53).

Materials

Each of 75 lists per participant was comprised of six 2-syllable nouns drawn randomly from the Toronto Word Pool (Friendly et al., 1982), displayed in capital letters one word at a time at the centre of the screen. Due to random shuffling of the original word pool anew for each participant, the probability for each word/serial position combination was equal.

Procedure

The experiment was run in PsychoJS, developed in combination with python code, using the PsychoPy3 library (Peirce et al., 2019) autotranslated to PsychoJS and run remotely through Pavlovia.org. Independent of grouping, each participant was asked to type the presented list in the forward or backward order (each word, itself, was to be typed forward). Each trial began with an asterisk in the centre of the screen, followed by the list of words from the originally shuffled pool presented sequentially one at a time in the centre of the screen.

The three conditions are depicted in Figure 2. Items were presented for 500 ms each. For the control and subjective chunking groups, 450-ms blank pauses were delivered between words. For the temporal grouping condition, there were 350-ms pauses within each group and an additional 600-ms pause in between groups for a total of 950 ms between items 3 and 4. Following the final 350-ms delay for the temporal grouping or 450-ms delay for both control and subjective chunking conditions, participants were asked to recall the studied list in a particular order at the centre of the screen. Recall was not time-limited. Participants were instructed to press ENTER following each recalled word. Only the single letter-string they were working on was displayed. They could edit the current word with the BACKSPACE key. However, once submitted with the ENTER keypress, the word disappeared from the screen and could not be edited further. Recall of the list ended after submitting 6 responses. Participants were also instructed to type "PASS" if they could not remember a particular word in a particular serial position. Only English alphabet letters were accepted, and the entered letters stayed on the screen until the ENTER key was pressed. Immediately after, participants could press ENTER to start another list. To reduce participants speeding through the task, the ENTER key (to submit the response) was only available once a minimum of three letters had been typed and those three letters could not be identical. Keyboard repeat was disabled. Age and gender were collected at the end of the session but not analyzed.

Data analyses

Data were analyzed in MATLAB (The Mathworks, Inc.) and JASP (JASP Team, 2023). Accuracy was scored based on a strict serial-position criterion; a word was correct only if it was recalled in the correct position and spelled correctly. Inter-response times were onset-to-onset (first letter) of typing each response for correct responses (the later item of the transition) only. Initiation times were computed from the cue to recall to the

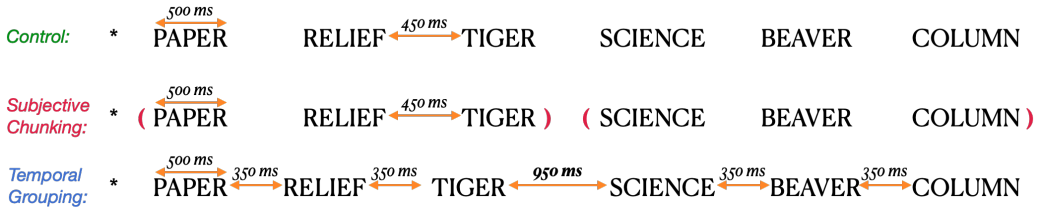

Figure 2

An illustration of the procedure of Experiment 1 with six sample words. Note that the total study time, from the onset of the first word until the presentation of the cue for recall was equated across all grouping conditions and recall directions. For each condition, half the participants were asked to recall the list in the same order (forward) as presented and half in the reverse order (backward).

first keypress, also for correct responses only. For Classical ANOVAs, the Greenhouse-Geisser correction was applied for violations of sphericity. Bayes factors were computed with JASP. BF_{10}, the ratio of evidence for the hypothesis versus the null, and $BF_{\text{inclusion}}$, the ratio of evidence for inclusion of an ANOVA effect versus omitting it (considering it null), are considered "some" evidence for the effect or null when $BF > 3:1$ or $< 1:3$, respectively, and within the $1:3$ to $3:1$ range, the result is considered underpowered and thus inconclusive (Kass & Raftery, 1995).

Results and discussion

First we report accuracy and inter-response time serial position curves, looking out for the characteristic scalloping (slight advantage for the first and last item within each group) and additional pause between groups, respectively, that others have reported with digit and letter stimuli (e.g., Farrell & Lewandowsky, 2004; Frankish, 1985; Hitch et al., 1996; Liu & Caplan, 2020; Ryan, 1969a, 1969b) and word stimuli (Spurgeon et al., 2015). Then we turn to our main question, and examine TEPs to test for all-or-none retrieval of groups. Finally, we report within-list intrusions to check for the presence of interpositions, items recalled from a different group but at the correct within-group position, a test of positional cueing (Henson, 1996).

Accuracy

Serial-position curves (Figure S1a,b; cf. data from Liu and Caplan (2020) in panels c,d) showed benefits of temporal grouping and subjective chunking in both recall directions and some of the classic scalloping effects. We asked whether the benefits of grouping or chunking would be a function of serial position or of output position; if the latter, as was found for consonant lists (Liu & Caplan, 2020), then the three-way interaction, Output Position×Condition×Direction should be non-significant. In a mixed ANOVA on accuracy, with design Output Position[6]×Condition[3]×Direction[2] (Table 1a), the three-way interaction was, indeed, a supported null ($p = 0.113$, $BF_{\text{inclusion}} = 0.108$), extending this same kind of result from nine-consonant lists (Liu & Caplan, 2020) to six-word lists. There

was a significant two-way interaction of Output Position×Condition. Both grouping strategies performed almost identically across all output positions, with significant improvement over the control group at certain output positions. Post-hoc pairwise t-tests found significant differences of temporal grouping versus control at positions three and six, $t(80) = -3.10$, $p = .003$; $t(80) = -2.41$, $p = .018$, respectively; and subjective chunking versus control at position three, $t(108) = -3.28$, $p = .001$, partly consistent with typical scalloping effects and partly replicating scalloping for temporally grouped word lists (Spurgeon et al., 2015).

Inter-response times

A between-group pause would be expected between recalled items 3 and 4 (Figure S2). Visual inspection suggests this may have occurred in the backward direction but only minimally in the forward direction. A mixed ANOVA (Table 1b) on mean inter-response time for correct recalls, with design Transition[5]×Condition[3]×Direction[2] produced mostly supported null effects, including for the three-way interaction; as with accuracy, any effects of grouping or chunking were equivalent when aligned by output position. The main effect of Direction indicates that backward recall was slower, on the whole, than forward recall. The interaction, Transition×Condition, was significant, although the supported-null Bayes Factors suggests that the effect was small in magnitude and should be interpreted with caution. The Transition×Direction interaction was significant and supported by a Bayes Factor above 3. To understand these, we conducted our six planned comparisons, using Tukey's correction for all pairwise comparisons, testing for a significant time increase during the transition from output position two-to-three to the transition from three-to-four, in each Condition×Direction sub-condition. All transitions were not significantly different ($p > 0.9$) in the forward direction and for the control group in the backward direction. Inter-response times were significantly slower in transition three-to-four than two-to-three for subjective chunking ($p < 0.001$) and temporal grouping ($p = 0.023$), respectively. The complexity of the stimuli and range of word length may have produced more variability in inter-response times than has been seen for consonants (Liu & Caplan, 2020). Analysis of the initiation times produced only supported null effects (Table S1).

Transitional-Error Probabilities

An ANOVA (Table 1c, plotted in Figure 3a,b) with design Transition[6]×Condition[3]×Direction[2] returned a supported null three-way interaction, suggesting that the significant interaction (with conclusive Bayes Factor), Transition×Condition, was itself no different for backward than forward recall. Aside from these quantitative analyses, it is clear from visual inspection that the telltale sign of all-or-none retrieval is not convincingly present. The expected stark drops in transitional error rates for the transitions $1 \rightarrow 2$, $2 \rightarrow 3$, $4 \rightarrow 5$ and $5 \rightarrow 6$ are actually not particularly stark, and in fact, sometimes go in the wrong direction. Participants tend to make more errors following recall of the first item of a putative chunk, not fewer. In the forward direction, second chunk, the error rate within-chunk (transition 4→5) is even greater than the error rate into the chunk (3 →4). That said, transitional error rates are generally greater in the control group than both temporal grouping and subjective chunking groups, which is what the all-or-none hypothesis would lead one to expect. However, the condition-differences are quite small compared to what even approximately all-or-none retrieval implies. Although

	Effect	F	df, error df	MSE	p	η_p^2	BF
a	Accuracy						
	Output Position	616.020	2.075, 549.787	0.045	<0.001	0.699	>1000
	Condition	2.740	2, 265	0.153	0.066	0.020	40.187
	Direction	0.399	1, 265	0.153	0.528	0.002	>1000
	Output Position×Condition	4.130	4.149, 549.787	0.045	0.002	0.030	152.643
	Output Position×Direction	12.820	2.075, 549.787	0.045	<0.001	0.046	>1000
	Condition×Direction	0.43	2, 265	0.153	0.651	0.003	0.289
	Output Position×Condition×Direction	1.864	4.149, 549.787	0.045	0.113	0.014	0.108
b	Inter-Response Times						
	Transition	27.378	2.875, 710.153	0.667	<0.001	0.100	>1000
	Condition	1.490	2, 247	7.914	0.227	0.012	0.182
	Direction	0.21	1, 247	7.914	0.018	0.022	3.555
	Transition×Condition	2.672	5.750, 710.153	0.667	0.016	0.021	0.134
	Transition×Direction	5.017	2.875, 710.153	0.667	0.002	0.020	5.230
	Condition×Direction	0.003	2, 247	7.914	0.316	0.009	0.216
	Transition×Condition×Direction	0.908	5.750, 710.153	0.667	0.486	0.007	0.003
c	Transitional-Error Probabilities						
	Transition	219.070	2.790, 719.748	0.048	<0.001	0.459	>1000
	Condition	2.789	2, 258	0.119	0.063	0.021	3.647
	Direction	0.21	1, 258	0.119	0.174	0.007	>1000
	Transition×Condition	2.672	5.750, 719.748	0.048	0.002	0.027	15.492
	Transition×Direction	5.017	2.875, 719.748	0.048	<0.001	0.063	>1000
	Condition×Direction	0.003	2, 258	0.119	0.853	0.001	0.138
	Transition×Condition×Direction	0.908	5.750, 719.748	0.048	0.469	0.007	0.007

Table 1

Experiment 1: ANOVAs on accuracy (a) with design Output Position[6]× Condition[Control,Subjective Chunking, Temporal Grouping]× Direction[Forward,Backward] and mean inter-response time (b) with design Transition[5]× Condition[Control,Subjective Chunking, Temporal Grouping]× Direction[Forward,Backward] and on Transitional-Error Probabilities (c) with design Transition[6]× Condition[Control,Subjective Chunking, Temporal Grouping]× Direction[Forward,Backward]. BF = $BF_{inclusion}$.

evidence of all-or-none retrieval is not a dominant characteristic of the temporal grouping or subjective chunking data, the relative difference compared to Control is broadly in the expected direction. Although we cannot conclude that all-or-none retrieval characterizes the error data as a whole, there might be a minority of participants or lists that would be accurately described as all-or-none.

Finally, consider order of report. It would be reasonable to argue that for these once-presented lists, chunks are only starting to form. In the forward direction, the first chunk exhibits a bit more of the expected all-or-none retrieval pattern than the second chunk. During study, it is thus possible that the first three words are more likely to be formed into a chunk than the last three words. However, the same pattern is obtained for backward recall— but as a function of output position, not serial position. The question then becomes, why is the first chunk *reported* more likely to be learned as a chunk than the first chunk *studied*. A limitation of this experiment is that recall direction was pre-cued (manipulated between subjects), so participants did have the freedom to potentially have prioritized the first studied chunk if they were in the forward direction, but the second studied chunk if they were in the backward direction. We are, however, quite a distance from the notion of all-or-none retrieval that we started with. With nearly the same design, Liu and Caplan (2020) found largely similar effects of temporal grouping as a function of output position when consonant lists were pre-cued (their Experiment 1) as post-cued (their Experiment 2). A more parsimonious view is that transitional-error probabilities are a function of output position and the lists are not yet formed into hierarchical memories.

Transitional-Error Probabilities of the consonant-list data. Finally, we wondered how the transitional-error probability pattern of the 9-consonant lists of Liu and Caplan (2020) would compare (data retrieved from https://osf.io/evmct). For both experiments, with direction manipulation between subjects (Experiment 1) or within-subjects, post-cued (Experiment 2), the all-or-none characteristic was visible for temporal-grouping participants but not for control participants (Figure 3c,d). Note, especially, that almost all within-chunk transitions had lower error rates for grouped than control participants, and many of these were significant. A notable exception is the first chunk retrieved in the forward direction in the pre-cued experiment (Figure 3c), but that might be obscured by a ceiling effect at those positions. The experiment with direction post-cued (Figure 3d), in fact, showed even more statistically robust differences between grouped versus control in the expected directions. This arguably aligns even better than the data from Johnson (1970) with the assumption of all-or-none retrieval.

Within-list errors

The presence of interpositions in within-list intrusions (Figure 4), presumed to indicate cueing with within-chunk position, would be elevated error rates from positions 4, 5, 6, 1, 2, and 3 (at positions 1–6, respectively), and especially, greater rates of such errors in the Subjective Chunking and Temporal Grouping data than Control. There are examples of small peaks from those interposition positions, and they are even the maximum source of within-list errors in the Forward, Temporal Grouping data at the first and last positions and in the Forward Subjective Chunking data (c) at output position 3. There is little trace of interpositions in the backward data. Rather, immediate adjacent transpositions are the most common source of within-list errors in all conditions. In sum, interpositions are, as

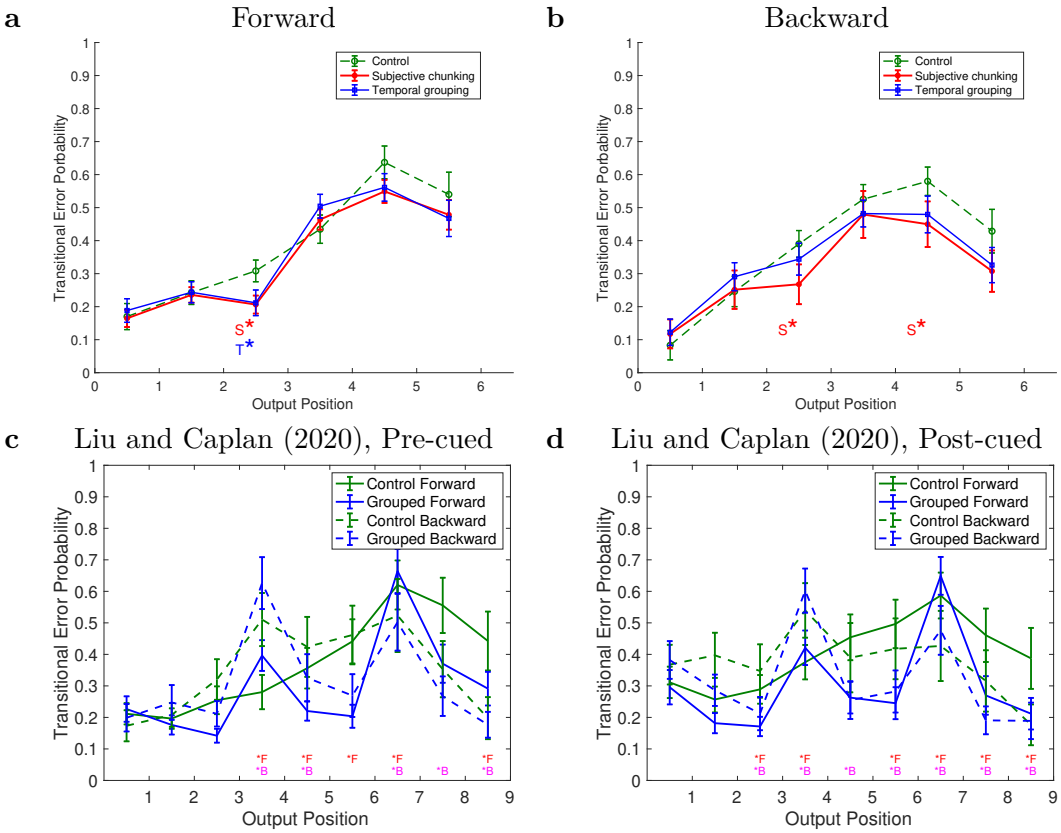

Figure 3

*Experiment 1: Transitional-Error Probabilities (probability of an error, given that the prior item was correctly recalled), in the forward (a) and backward (b) recall conditions. Participants with one or more missing values were omitted (Control Forward: 2 of 42, Backward: 0 of 43; Subjective Chunking Forward: 2 of 68; Backward: 0 of 31; Temporal Grouping Forward: 0 of 40, Backward: 3 of 47). For (a) and (b), asterisks denote significant differences between Control and Subjective Chunking (*S) or Temporal Grouping (*T) conditions, respectively. The 9-consonant list data from Liu and Caplan (2020), where recall direction was pre-cued in Experiment 1 (c) and post-cued in Experiment 2 (d). For (c) and (d), asterisks denote significant differences between grouped and control for forward (*F) or backward (*B) direction, respectively. Points are plotted between the two output positions comprising the transition. Error bars plot 95% confidence intervals based on standard error of the mean.* ***The first point plots initiation error probability.***

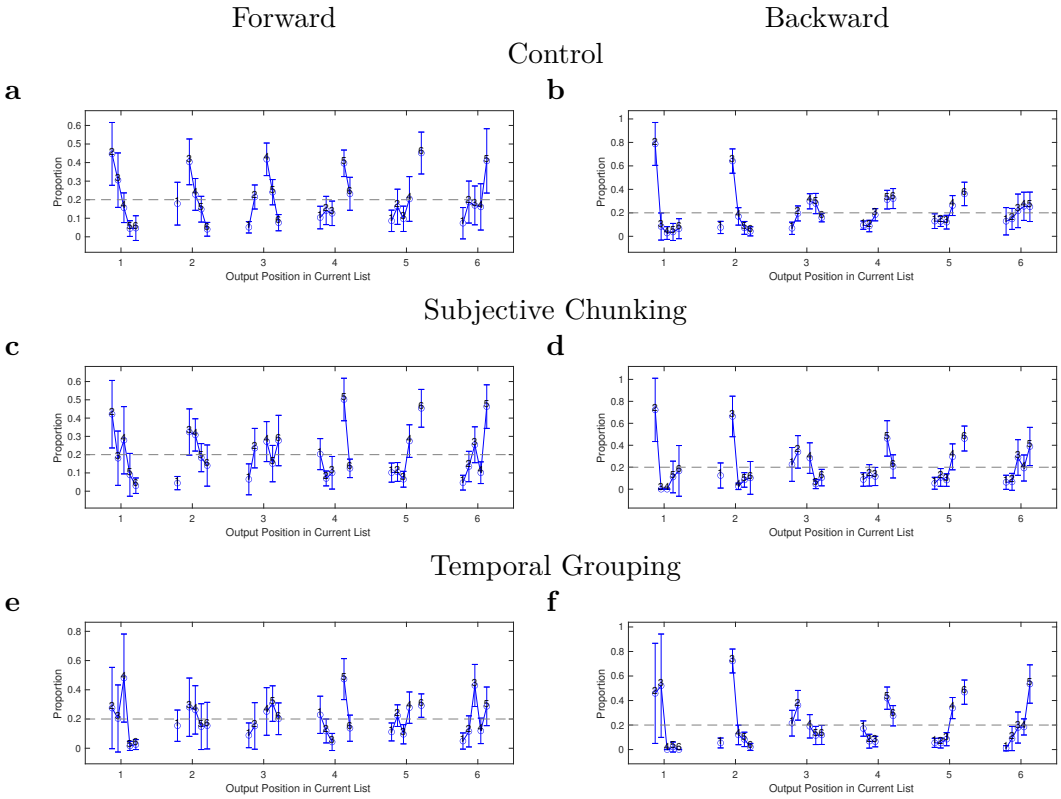

Figure 4

Experiment 1: Rates of within-list intrusions as functions of output position in the current list (x axis) and position within the previous list (denoted by numbers next to the means). These values were normalized to express them as proportions of within-list intrusions at a given output position. Participants with ceiling or floor performance were excluded. Error bars are 95% confidence intervals based on standard error of the mean.

previously reported, more common in the temporal grouping and subjective chunking conditions in the forward, but not backward direction. However, different than what has been previously presumed, interpositions are not made at all output positions. The lack of coupling of interpositions with scalloped accuracy serial position effects (Figure S1) converges with Liu and Caplan's (2020) findings with consonant lists.

This poses a problem for accounts of chunking or grouping that assume retrieval is carried out in part by cueing with the within-chunk position. Such a mechanism would predict interpositions at all output positions and also in backward recall. To check our intuition, we ran the favoured version of SIMPLE using the best-fitting parameter set for each participant from Liu and Caplan (2020) and plotted the resulting within-list intrusion pattern produced by the model. Figure 5 shows that interpositions were clearly present and quite prevalent at all output positions, for the Forward, Grouped condition. The same characteristic was observed for Control and Backward recall directions (not shown).

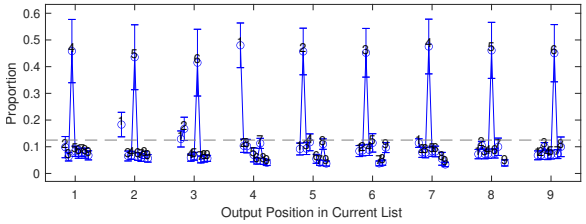

Figure 5

Simulation of SIMPLE: Rates of within-list intrusions produced by the model fits of the Liu and Caplan (2020) data, Forward Grouped condition. Each participant's best-fitting parameter set was executed, taking as inputs the participant's mean response time, as done in Liu and Caplan (2020). Within-list intrusions were then computed as in Figure 4, as functions of output position in the current list (x axis) and position within the previous list (denoted by numbers next to the means) and again normalizing at each output position separately. Interpositions are responses at ± 3 *or* ± 6 *positions from the current output position. Error bars plot 95% confidence intervals based on standard error of the mean.*

Summary of Experiment 1

Scalloping of serial-position curves was observed for both temporal grouping and subjective chunking conditions, showing consistency with prior digit and letter studies (e.g., Farrell & Lewandowsky, 2004; Frankish, 1985; Hitch et al., 1996; Liu & Caplan, 2020; Ryan, 1969a, 1969b) as well as temporal grouping of word lists (Spurgeon et al., 2015). Effects of both temporal grouping and subjective chunking were largely functions of output position, not serial position, replicating the pattern found by Liu and Caplan (2020) with lists of words. This suggests both grouping and chunking with the current procedures primarily affect processes that unfold over the course of the recall sequence rather than acting primarily during the study phase.

Regarding our main question of interest, the TEP pattern, no indication of all-or-none retrieval (i.e., a spike in TEPs between chunks accompanied by a drop in TEPs within chunk) was evident and to some degree, the results exhibited the opposite pattern. In contrast, and also as proof-of-principle that the TEP function can, in fact, produce an spike-like pattern, we found compelling evidence in favour of the all-or-none retrieval pattern in the previously published consonant data for the temporally grouped condition only. This is consistent with the idea that letter lists are easily recoded. In fact, given that participants typed their responses, it is also conceivable that they were recoding not only based on words or acronyms but also based on the spatial pattern of the letters on the keyboard, itself.[1] The pronounced TEP spikes were seen in backward as well as forward recall directions. This further suggests that recoded subsequences are straight-forward to unpack in either direction, the kind of flexibility suggested by previous studies (e.g., Cowan et al., 2002; Ward & Tan, 2019; Watkins & Bloom, 1999). Regarding mechanisms of subdivision, interpositions are predicted if within-group position is a cue during recall, as demonstrated by an implementation of the model SIMPLE. However, interpositions were also scarce, questioning the use of within-group position as a retrieval cue.

[1] We thank Geoff Ward for this idea.

In sum, where recoding is plausible, with lists of consonants, we saw striking TEP spikes suggestive of all-or-none retrieval. But for word lists studied once, no clear trace of all-or-none retrieval was found.

Experiment 2

Experiment 1 yielded little evidence of all-or-none retrieval of chunks that participants would have had to construct on the fly. Returning to the song-list thought experiment in the Introduction, we wondered if novel lists studied once might produce all-or-none retrieval if they were constructed from previously trained chunks, adapting the approaches of Chen and Cowan (2005) and Thalmann et al. (2019). In those procedures, items within chunks were treated differently; the first item was provided as the retrieval cue during training, whereas the remaining items were practised by recalling them. We designed a training procedure wherein all items within each chunk would receive recall-practice.

We considered two approaches to pre-training four three-word chunks, both with multiple iterations of study followed by serial-recall: study/test of each chunk alone or study/test of a continuous list where participants had to demarcate the chunk boundaries. First, each three-word chunk could be studied, followed with a distractor and then recalled right away before proceeding to the next chunk. We refer to this training set as 4×3. Because of the short list length, we expected accuracy to be high, even with a short end-of-list distractor. Because those chunks were to be combined into new probe lists after a much longer delay, accuracy during training in the 4×3 condition very likely overestimates the participant's degree of learning. Second, we were curious if the all-or-none characteristic might be undermined if participants had to extract chunks from a continuous serial list, where in each training iteration, each "chunk" was always preceded and followed by the same other words. In this 1×12 training procedure, the four chunks were not demarcated as such but studied in a sequence of 12 words, with instructions to the participant to subdivide the list into groups of three words each. We speculated that participants may learn inter-chunk associations, which might disrupt participants' representations of chunks as isolated subsequences in memory (however, no substantial differences between probe lists derived from the 4×3 versus 1×12 training sets were found). Because of the longer list length, during training, there was no end-of-list distractor; participants attempted forward serial recall of the twelve words after the last item was presented. Because of the longer list length, accuracy during training may be more reflective of degree of learning of the chunks than for the 4×3 training.

In detail: in our procedure, in the first training set, 4×3, four chunks of three words each were trained over three study–test serial recall cycles, with an interpolated arithmetic distractor, where chunk order was randomly shuffled from one cycle to the next. The distractor was included in an attempt to bring accuracy down from ceiling, to give us a sense of the degree of learning of the pre-trained chunks, although accuracy still ended up quite high (Figure S6).

In the second training set, 1×12, chunks were trained over the course of three study–test cycles of a 12-item list, presented in the same order on each cycle. For this set, participants were instructed to subdivide the list into 4 groups of 3 words each, comparable to the Subjective Chunking condition of Experiment 1.

Finally, probe lists were constructed from random sequences of those pre-trained chunks to test whether such pre-trained chunks would be retrieved all-or-none. Probe lists alternated being derived from the 4×3 or the 1×12 set.

We expected the 4×3 training would produce more isolated chunks, as items studied close in time would have been random. For the 1×12 training, we speculated that the consistent chunk-order would make it difficult for participants to cleanly isolate one chunk from another chunk, despite being instructed to subdivide the list into chunks. As it turned out, the two training procedures produced very similar behavioural patterns.

Methods

Methods were the same as in Experiment 1 wherever possible.

Participants

A total of 60 participants were recruited from Prolific website (www.prolific.co) and completed the experiment on the Pavlovia website (www.pavlovia.org) in exchange for GBP £8.5. We targeted $N = 60$ as a rough mid-point of the final sample sizes obtained in Experiment 1. We planned to run additional participants in sets of 10/group (training order) in an effort to obtain conclusive Bayes Factors for our effect of interest (the three-way interaction Transition×Training set×Direction on Transitional-Error Probabilities) but this turned out not to be necessary. A robustness check of the effect size as suggested by R. B. Anderson et al. (2022), plotted in Figure S11b, shows that the η_p^2 was fairly stable toward the end of data-collection. Participants were assigned to one of two training orders (which training procedure was performed first) alternating in order of recruitment. After excluding 9 participants with floor/ceiling accuracy (average fewer than 1 word recalled per list or more than more than 8 words recalled per list, respectively) accuracy, the final numbers of recruited participants were: first training-order: 27 (out of 32), second training-order: 24 (out of 28).

Materials

Stimuli were the same as in Experiment 1. The lists were constructed differently.

Procedure

Every list presentation began with an asterisk in the centre of the screen. The experiment consisted of five phases: practice, baseline, chunk training (4×3 and 1×12, order varied across participants) and probe-list.

Practice phase

Participants were given instructions and two lists of six words to practice forward and backward recall using the same procedures as in Experiment 1, except each word was presented for 2000 ms with a 350-ms blank inter-stimulus interval. Because this may have been the participants' first encounter with forward and backward serial recall, we do not report the data from this phase.

Baseline phase

Baseline serial-recall performance in both directions was assessed with four lists of six newly sampled words each, with identical procedures as in the practice phase. Recall direction was post-cued, in fixed order for all participants: Forward, Backward, Forward, Backward.

Chunk training phase

In the third and fourth phases, each participant practised two separate sets of twelve words over three consecutive cycles, recalling them in the forward direction. Procedures for list presentation and serial recall were identical to the practice and baseline phases, although list length differed as follows.

The 4×3 set was comprised of twelve words (not previously used for the participant), divided into four groups of three words each, and these groups were subsequently used to construct probe lists (next phase). After studying each group sequentially, participants answered three arithmetic equations (distractor task) and then recalled the studied group in forward order only.

The 1×12 set was a new list of twelve words not previously used. These twelve words were presented sequentially, followed by forward serial recall of the entire list with no intervening distractor task. Participants were instructed to subdivide the twelve-word list into four groups of three words each, and these subjective chunks were subsequently used to construct probe lists (next phase).

For each training set, the set was studied entirely, three times successively. On each iteration, the 4×3 set was studied in a new group-order, whereas the 1×12 set was always studied in the same order. Half the participants were trained on the 4×3 set followed by the 1×12 set, and the other half were trained in the opposite order in alternating order of recruitment.

Probe-list phase

In each of 28 final probe-list cycles, three groups from one of the training sets were concatenated to form a new nine-word list. Probe lists alternated being derived from each of the training sets, starting with the first training set studied for a given participant. Words were presented for 1500 ms each, with a 350-ms blank inter-pair interval. The presentation rate was sped up during the piloting stage, due to concerns that accuracy might be too close to ceiling. Recall was typed as before, in forward or backward order, with the direction of recall counter-balanced within each source set (14 cycles each). Recall of a list ended after submitting nine responses.

Age, gender, sex, and aphantasia self-report were collected at the end of the session but not reported here. Age and gender were not analyzed and aphantasia rates were too low to be interpretable.

Data analyses

Data were analyzed using the same approach as in Experiment 1. Due to the smaller number of (probe) lists in this experiment compared to Experiment 1, restricting

inter-response time analyses to correct recalls resulted in too many missing values, so we report inter-response times and initiation times regardless of accuracy.

Results and discussion

We report data from the baseline serial-recall phase in the Supplementary Materials. Standard serial-position effects, including the predominant dependence on output position, were found for these six-word lists in the absence of temporal grouping or instructions to chunk. Transitional-error probabilities showed no evidence of all-or-none retrieval in these six-word, ungrouped lists, providing some continuity with the ungrouped condition of Experiment 1.

Next, we report learning-curve data for the training phase and then serial-position curves for the probe-list phase. We then turn to our main goal, to test for all-or-none retrieval of pre-trained chunks assembled into new lists (the probe lists) and finally, we examine within-list intrusions for the probe-list phase, to determine the relationship between interpositions and all-or-none retrieval.

Chunk-training

To understand the effects of grouping on the training efficiency, Strict recall accuracy for both training sets were plotted as a function of cycle (Figure S6). Participants were able to correctly recall around 90% of the 4×3 set by the second cycle. However, they correctly recalled only around 50% of the 1×12 set after three cycles. A repeated-measures ANOVA on serial-recall accuracy with design Training Set[4×3, 1×12]×List Number[3] produced highly significant main effects and interactions ($p < 0.001$). We were most concerned with whether there was any change in accuracy over the course of training. Simple effects confirmed a significant main effect of List Number for both training sets. Post-hoc pairwise comparisons, Bonferroni-corrected, revealed that all pairwise differences were significant ($p < 0.01$) aside from list 2 versus list 3 for the 4×3 set ($p > 0.5$). In other words, there is evidence that participants improved over the course of training from the first to the second list, but from the second to third list, the 4×3 set might be subject to a ceiling effect.

Probe-list Accuracy and Inter-response times

Serial-position curves (Figure S7, ANOVA reported in Table 2) show largely the same effects for probe lists derived from the two training sets. Although the interaction, Output Position×Training List was significant, the low Bayes Factor and η_p^2 suggest the interaction is small.

Mean inter-response times, regardless of accuracy (Figure S8, ANOVA reported in Table 2) show evidence of long pauses between chunk transitions across both recall directions, except for the last chunk in forward direction ($p<0.05$).

Thus, the characteristic pause prior to recall of a chunk was very prominent in recall of the probe lists. This reinforces the idea that the probe lists were being treated as subdivided. Analysis of the initiation times produced only supported null effects or nearly so for the main effect of Training set (Table S2). It must be noted that unlike Experiment 1 and due to the low number of overall submissions, none of the incorrect responses were excluded.

	Effect	F	df, error df	MSE	p	η_p^2	BF
a	Accuracy						
	Output Position	81.789	2.450, 122.497	0.162	<0.001	0.621	>1000
	Training set	0.260	1, 50	0.107	0.612	0.005	0.076
	Direction	2.190	1, 50	0.381	0.145	0.042	0.730
	Output Position×Training set	2.752	5.316, 265.823	0.024	0.017	0.052	0.077
	Output Position×Direction	2.602	2.546, 127.285	0.162	0.064	0.049	1.888
	Training set×Direction	0.082	1, 50	0.054	0.775	0.002	0.035
	Output Position×Training set×Direction	1.215	4.823, 241.172	0.029	0.303	0.024	0.001
b	Inter-Response Times						
	Transition	23.397	3.080, 153.988	8.601	<0.001	0.319	>1000
	Training set	6.155	1, 50	1.076	0.017	0.110	0.165
	Direction	16.548	1, 50	1.537	<0.001	0.249	20.312
	Transition×Training set	1.254	5.253, 262.675	1.477	0.283	0.024	0.008
	Transition×Direction	2.975	4.118, 205.919	2.793	0.019	0.056	5.141
	Training set×Direction	0.119	1, 50	0.847	0.732	0.002	0.067
	Transition×Training set×Direction	0.335	4.558, 227.918	1.676	0.876	0.007	<0.001
c	Transitional-Error Probabilities						
	Transition	10.573	5.230, 120.281	0.058	<0.001	0.315	>1000
	Training set	0.400	1.000, 23.000	0.062	0.533	0.017	0.064
	Direction	0.283	1.000, 23.000	0.037	0.600	0.012	0.057
	Transition×Training set	0.340	4.907, 112.857	0.048	0.885	0.015	0.002
	Transition×Direction	1.462	4.795, 110.289	0.049	0.210	0.060	0.022
	Training set×Direction	0.360	1.000, 23.000	0.014	0.555	0.015	0.007
	Transition×Training set×Direction	0.739	4.614, 106.124	0.046	0.585	0.031	<0.001

Table 2

Experiment 2: ANOVAs on accuracy (a) with design Output Position[9]×Training set[4×3, 1×12]×Direction[Forward, Backward] and mean inter-response time for correct responses (b) with design Transition[8, excluding the initiation times]×Training set[4×3, 1×12]×Direction[Forward, Backward] and on Transitional-Error Probabilities (c) with design Transition[9]×Training set[4×3, 1×12]×Direction[Forward, Backward]. BF = $BF_{inclusion}$.

Probe-list Transitional-Error Probabilities

With only floor and ceiling participants removed from the analyses, some support for the presence of TEP spikes was found (Supplementary Materials, Figure S9). However, TEPs are computed only following correct recalls; for participants with no correct recalls at a given output position, TEPs at the subsequent position are undefined. Because different participants had different missing values, eliminating missing values cell-wise may be misleading. We report TEPs for the subset of participants with no missing values, in Figure 6 and ANOVA in Table 2. Although not clear-cut, the TEP spike pattern can be seen to emerge in the probe lists, particularly when one compares with Experiment 1 (Figure S1) and the baseline lists of Experiment 2 (Figure S3). In the forward direction, TEPs drop considerably from onset of the chunk to transitions within-chunk during recall of the first and last chunks and for the middle chunk as well when derived from the 1×12 set but not the 4×3 set. In the backward direction, the first recalled chunk shows no spike, but that is because initiation of recall is quite accurate in the backward direction; transitions within-chunk are extremely accurate for the first-recalled chunk. The second and third recalled chunk show clear drops in error rate from initiation to within-chunk transitions, and similarly for both training sets. All interactions were non-significant, supported nulls. This suggests that the observed effects are driven by output order (the Transition factor) rather than serial position, pointing to processes acting primarily during the course of recall. The one significant effect was the main effect of Transition. Bonferroni-corrected pairwise post-hoc comparisons revealed transition 3→4 and 6→7 had a significantly ($p < 0.05$) greater error rate than all other transitions apart from each other. All other comparisons were non-significant. This confirms the relatively lower error rate within-chunk than initiating the chunk (transitions 3→4 and 6→7).

To appreciate the effect of excluding participants, in Figure S10 we report TEPs for each of the three groups at a time, and for each group, we include all participants with no missing values for that chunk. This includes more participants, and qualitatively resembles Figure 6.

Within-list errors

Our goal, following Henson (1998) and Henson (1996), was to seek evidence that participant use within-chunk position as a retrieval cue, which would materialize in high rates of interpositions. Interpositions were not particularly reliably found for word lists in Experiment 1, nor for consonant lists even when TEPs suggested all-or-none retrieval (Figure 4), which may be conducive to recoding. Depending on how recoding works, it is possible that recoding is either compatible or incompatible with cueing with within-group position. Here we ask whether interpositions are common when recall appears all-or-none (TEP analyses reported in the previous section), in lists of words that are unlikely to be conducive to recoding. We report within-list intrusions for the subset of participants who had no missing values in the TEP analyses, and further excluded participants with fewer than two within-list intrusions.

Figure 7a,b reveals high rates of within-list intrusions at ±3 and ±6 lags (interpositions), along with the expected high frequency of adjacent intrusions, in both forward and backward recall directions. However, these could reflect retrieval of the entire chunk in the

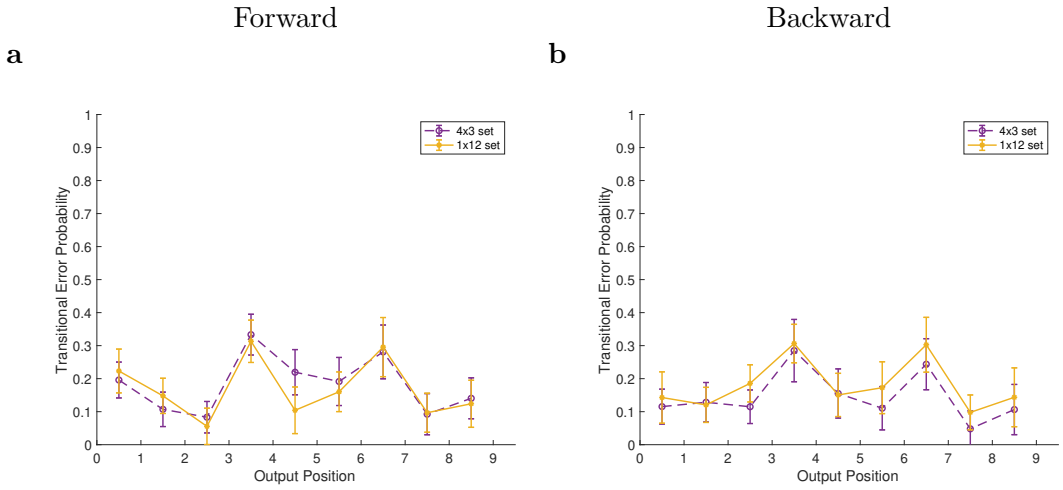

Figure 6

Experiment 2: Transitional-Error Probabilities as a function of transition for forward (a; $N = 35$ and $N = 32$ for the 4×3 and 1×12 sets, respectively) and backward (b; $N = 26$ and $N = 31$ across 4×3 and 1×12 sets, respectively) recall directions. Participants with one or more missing values were omitted. Error bars plot 95% confidence intervals based on standard error of the mean. ***The first point plots initiation error probability.***

wrong position, as noted by Lee and Estes (1981). This is plausible given that all-or-none retrieval was fairly well supported for this particular data set. Such whole-chunk displacement obscures any interpositions that might plausibly arise from using within-chunk position as a retrieval cue. To eliminate such cases, we next restricted the analysis of within-list intrusions to responses following correct responses (Figure 7c,d). Note that output positions 1, 4 and 7 could still reflect retrieval of the wrong chunk *in toto*, but the remaining output positions would not be subject to this ambiguity. The middle and final output positions do not show a consistent pattern of interpositions, and are largely overshadowed by peak intrusion rates at adjacent positions or the final list item (9). The main exception where interposition rates are high is in the forward direction, output position 9, where item 6 is frequently recalled.

In sum, when participants recall lists composed of previously learnt chunks, and recall appears approximately all-or-none, interpositions appear due to the trivial account (Farrell, 2012; Henson, 1996; Lee & Estes, 1981) that the entire chunk is recalled out of order, but when such cases are eliminated, little trace of interpositions is left. This suggests that within-chunk position is not a major retrieval cue used in all-or-none retrieval of chunks.

Recall of the omitted chunk

To make the probe lists more challenging, they were composed of three out of the four pre-trained chunks within a training set. This was meant to reduce the incentive for a participant to recall all chunks of a set and more or less guess at the chunk order; recall of the omitted chunk would be incorrect, and because of the shorter list length (9 words as opposed to the full 12-word set), recalling an omitted chunk would potentially displace

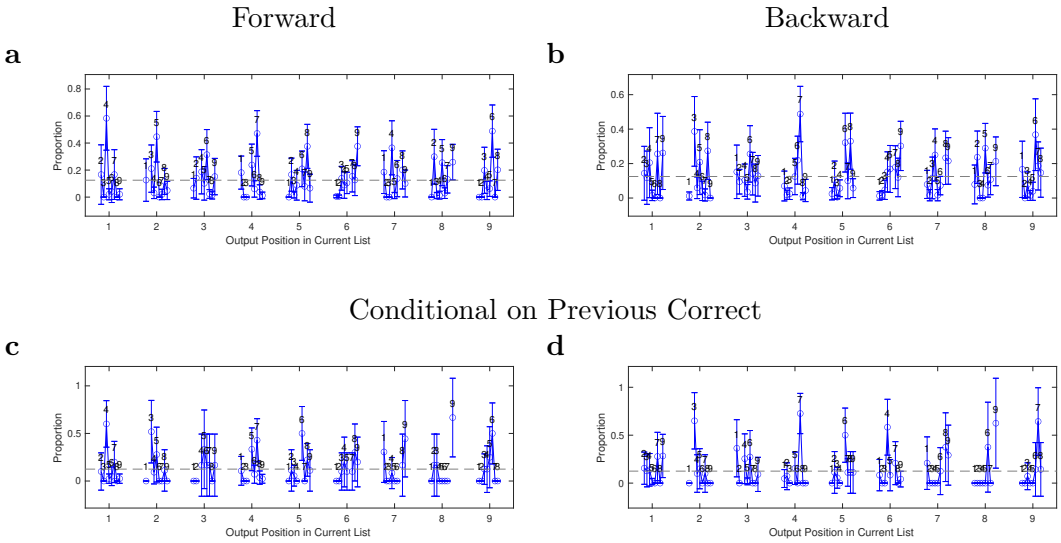

Figure 7

Experiment 2: Rates of within-list intrusions as functions of output position in the current list (x axis) and position within the previous list (denoted by numbers next to the means). These values were normalized to express them as proportions of within-list intrusions at a given output position. In (a) and (b), all within-list intrusions were included in the analyses. In (c) and (d), only within-list intrusions following correct recalls were included, to eliminate instances of recalling the entire chunk out of order. Participants with ceiling or floor performance were excluded, as were participants with missing TEP values and participants with fewer than two valid within-list intrusions in each respective analysis, leaving $N = 22$ participants in (a) and (b) and $N = 20$ in (c) and (d). Error bars are 95% confidence intervals based on standard error of the mean.

three correct responses. We were curious whether participants ever recalled items from the omitted chunk, and if so, how frequently was the chunk intact— recalled all-or-none.

Recalls of any word from the omitted chunk were already rare; collapsing across all participants, out of 714 total lists, one, two or three of the omitted-chunk words were recalled 33, 6 and 6 times, respectively for the 4×3 set and 27, 7 and 4 times for the 1×12 set. Recalls of 2 or 3 omitted-chunk words were thus more frequent than expected by chance (i.e., the product of the probability of a single omitted word recalled). Full details of all ten examples of all omitted words recalled are reported in Table S3. In these rare instances, whenever all three items were recalled, they were recalled a) in the correct order (with respect to recall direction) and b) aligned with the chunk-structure of the list. This suggests that when people recall a chunk in error, they retrieve it all-or-none, far more often than by chance. This is because the probability of reconstructing the correct order, alone, assuming the three items were planned for recall, is $((1/3)(1/2)(1) = 1/6 = 0.167)$. Obtaining that result 10 out of 10 times has probability $(1/6)^{10} = 10^{-7}$). Because these were such rare events, this null-hypothesis probability should be viewed with a great deal of caution. We report it only as anecdotal evidence that all-or-none retrieval, even of a

(pre-trained) chunk that was not part of the list, occurs occasionally.

General Discussion

Building on suggestive evidence of all-or-none retrieval reported by Johnson (1969, 1970) for lists of letters, our analyses of single-trial consonant-list data (Liu & Caplan, 2020) confirmed spikes in the transitional-error probability analysis consistent with all-or-none retrieval when letter-lists are grouped (spatially by Johnson or temporally by Liu and Caplan) but not when subdivision is not signalled to the participant (control condition of Liu and Caplan, 2020). Backward recall showed the same overall effects. All-or-none retrieval of letter-lists might be due to recoding (Ericsson et al., 1980; Miller, 1956), and in control lists, possibly the TEP effects are obscured by different participants opting for different recoding patterns as functions of serial position or even varying their recoding boundaries across lists. The striking reversal of the TEP pattern in backward recall (i.e., similarity to forward recall as a function of output position) suggests that recoded subsequences are straight-forward to "unpack" and reverse when required.

The two new experiments we report here examined word lists, which we thought would be more resilient to recoding. In Experiment 1, once-studied lists of words, although they did produce some scalloping in accuracy serial-position curves and evidence of long pauses between chunks in the backward but not forward direction, exhibited no clear evidence of all-or-none retrieval in TEP plots and to a large degree, transitional-error probabilities ran in the opposite direction than predicted. When chunks were pre-trained in Experiment 2, lists derived from those chunks did show compelling drops in transitional-error probabilities following the first recall of a chunk. In rare cases when items from the omitted chunk were recalled, they were far more likely than chance to be retrieved all together, in correct sequence (and correctly reversed in the case of backward recall) and even correctly aligned with chunk boundaries. The striking dependence on output position rather than serial position (backward recall compared to forward recall) suggests that all-or-none retrieval effects occur primarily during the recall sequence, and are not inevitably set up during the encoding phase, although some awareness of the presence of the pre-trained chunks surely must be critical during the study phase.

Finally, interpositions were scarce except when trivially observed due to displacement of entire chunks in Experiment 2, suggesting that retrieval using within-chunk position as a cue is rare, and perhaps even more rare in the regime in which retrieval resembles all-or-none.

In sum, temporal grouping and subjective chunking instructions do not appear to induce hierarchical encoding of lists in memory or all-or-none retrieval following a single study trial. Rather, with only a single study trial, subdivision may encourage recoding strategies when lists afford it, such as lists of letters or digits, and other mechanisms such as positional distinctiveness or inter-item associative strength for materials that cannot easily be recoded. With only a small amount of training on chunks as chunks, either by studying and recalling each chunk in isolation of the other chunks or by subjectively instructed chunk boundaries imposed on a single study sequence, lists composed of pre-trained chunks exhibit the signs of all-or-none retrieval. A parsimonious conclusion is that memory of a list is not inherently hierarchical but a non-hierarchical mechanism can be made to emulate hierarchically organized information over a few exposures.

Do the data demand a hierarchical memory model?

Most models of serial recall are not hierarchical. The best exception is Farrell (2012), whose model assumes a three-level hierarchy very much in line with the theory verbally described by Johnson (1969), where a list is composed of control elements which in turn are associated with items. This model successfully fit the data on temporal grouping of word lists (Spurgeon et al., 2015). Farrell explicitly assumed and emphasized that once its control element is retrieved, the constituent items are retrieved in forward order. The strict forward directionality is at odds with the near-mirror-image effects of backward versus forward recall that we found in both experiments as well as by Liu and Caplan (2020). However, it would seem a modest modification to assume that when instructed to recall backward, the model could cue with within-group position in reverse order and be able to handle the backward recall data. To our knowledge, Farrell's model has never been tested on TEPs, but because of the explicit sequential constraint, that the group's control element must be retrieved before the constituent items are retrieved, it seems quite plausible that the model would produce TEP spikes, lower error in transitions within-group than initiating a group. Farrell's model, because it assumed the only mechanism for grouping or chunking was the hierarchical structure of memory, would thus miss the TEP pattern (lack of clear evidence of all-or-none retrieval) in once-studied lists (Experiment 1 here as well as the control conditions of Liu and Caplan, 2020). It might still be a good account of the TEP pattern we found when word-chunks were pre-trained (Experiment 2) and letter-chunks induced by temporal grouping (Liu & Caplan, 2020) or spatial cues (Johnson, 1969, 1970). Farrell's model would also presumably over-predict the interposition rate and consistency across output positions compared to what we observed in both experiments and in the consonant-lists of Liu and Caplan (2020). We therefore suggest that Farrell's model may be overly complicated to explain effects of temporal grouping and subjective chunking in a completely novel-ordered, once-studied list, but its hierarchical design principles could provide useful clues to guide the extension of a non-hierarchical model to produce hierarchical-like performance in certain conditions, including words, themselves, construed as highly overlearned chunks as well as the song-"chunk" thought-experiment described in the introduction.

Cueing with within-chunk position

Positional-coding models have seen major success in explaining numerous findings in immediate serial recall. Such models associate each item with a positional or relative-order code. Those position codes are used as retrieval cues, one after the other, during serial recall. For grouped lists, it is assumed that both list position and within-group position are used as retrieval cues (e.g., Brown et al., 2007; Brown et al., 2000; Burgess & Hitch, 1999) or chunk position along with within-group position in the case of the Start-End Model (Henson, 1998). These model accounts were largely inspired by data from temporally grouped lists, where items (nearly always digits or letters) are presented sequentially, with an extra pause between groups (Ryan, 1969a, 1969b). Liu and Caplan (2020) found that such a model produced good fits of their temporally grouped consonant-list data. The effects of temporal grouping were largely a function of output position, not serial position, suggesting that the benefits of grouping unfold during the course of the recall sequence, hinting that the characteristic scalloping of the serial-position curve may be more related

to retrieval processes than to the way in which memories are stored. An adaptation of Scale-Invariant Memory, Perception, and Learning (SIMPLE; Brown et al., 2007), with the assumption of two-level positional cueing fit the data well. To achieve a good fit, it was also necessary to include repetition suppression, where items are removed from the pool of response candidates once recalled (Duncan & Lewandowsky, 2005).

The two-level positional cueing account of temporal grouping effects, which was a good fit to the data of Liu and Caplan (2020), leads one to predict a high rate of interpositions (Farrell & Lewandowsky, 2004; Henson, 1996; Hitch et al., 1996; Ryan, 1969a), where an item is displaced, but to the same relative position within a different chunk (we confirmed this in a simulation of SIMPLE that was previously fit to the Liu and Caplan data; Figure 5). Liu and Caplan (2020) found that proportions of errors that were interpositions were greater for grouped than for ungrouped lists, but only in the forward direction. Interposition rates did not covary with overall accuracy due to temporal grouping, weakening the support for positional-cueing as the correct account of temporal-grouping effects. If, as we suggested, recoding explains the effects of temporal grouping on lists of consonants (also in Johnson's experiments), perhaps interpositions are not common when recoded chunks are unpacked.

In our word lists in Experiment 1, interpositions were not particularly prevalent. In Experiment 2, when support for all-or-none retrieval was found in the TEP data, interpositions were also not clearly prevalent once displacements of entire chunks were ruled out, suggesting that when chunks are retrieved all-or-none, within-chunk position is also not used as a retrieval cue.

Taken together, interpositions (at all output positions) may indeed be evidence of the use of within-chunk position as a retrieval cue. This suggests boundary conditions: within-chunk position may not be used in retrieval of recoded chunks (such as in lists of digits or letters) or in retrieval of previously learnt chunks. Within-group position may be used as a retrieval cue, not when chunks are present, but when they are absent and within-group position is readily available to the participant.

Backward recall points to mechanisms operating during recall

Despite numerous interesting dissociations, serial-position effects largely reverse in backward versus forward serial recall (J. R. Anderson et al., 1998; Bireta et al., 2010; Cowan et al., 1992; Farrand & Jones, 1996; Guérard & Saint-Aubin, 2012; Guérard et al., 2012; Guitard & Saint-Aubin, 2021; Haberlandt et al., 2005; Hulme et al., 1997; Li et al., 2010; Li & Lewandowsky, 1993, 1995; Liu & Caplan, 2020; Madigan, 1971; Manning, 1982; Manning & Pacifici, 1983; Ritchie et al., 2015; St. Clair-Thompson & Allen, 2013; Thomas et al., 2003), indicating that output order is the largest factor determining recall probability of a given list-item. This is echoed in comparisons of forward serial recall to free recall (Grenfell-Essam & Ward, 2012; Tan et al., 2016) and data from serial recall starting from different starting points within a list (Cowan et al., 2002). This characteristic was largely replicated in both of the word-list experiments reported here. Moreover, the effects of chunking and grouping manipulations did not interact with recall direction. This points compellingly to the idea that effects of grouping and chunking predominantly influence the recall process rather than the encoding process. This does not rule out effects during study, and for example, recoding would seem to demand some cognitive activity during the study

phase. Rather, the effect on accuracy must influence the way in which each successive item is retrieved (see the notion of compression, well explained and reviewed by Norris and Kalm, 2021). Processes affected could include reducing or avoiding output interference, repetition suppression (which was the key to obtaining good fits of SIMPLE to the data in Liu and Caplan, 2020) and possibly, speeding recall time with the effect of reducing the effective retention interval (study–test time). All-or-none retrieval, when it occurs, may determine recall probability not only within a chunk, but also for subsequent items by reducing output interference or simply increasing accuracy to reduce the pool of response candidates or reduce time to recall.

Several proposals have been made that one cause of the advantage often seen for temporally grouped or chunked lists is that the grouping frees up cognitive resources such as, working memory or frees participants to perform other activities during pauses that benefit memory such as rehearsal (Barrouillet et al., 2004; Kowialiewski et al., 2020, 2022; Mızrak & Oberauer, 2021; Oberauer, 2003, 2022; Popov & Reder, 2020; Thalmann et al., 2019). When effects are quite focal and specific to specific serial positions (Liu & Caplan, 2020), such accounts would seem plausible. However, when the effects reverse in backward serial recall, as they did in Liu and Caplan (2020) and do again in both experiments reported here, the study phase becomes less credible as the locus of the effects. Effects of chunking and grouping that are primarily functions of output order rather than serial position point to processes that unfold over the course of recall, such as output interference, output encoding and repetition suppression, all of which are compatible with the pre-existing accounts such as all-or-none retrieval and positional cueing. At least in paradigms that show such reversal (functions of output position but not serial position), resource-based accounts would seem largely ruled out. The resource-based ideas might, however, be reformulated in terms of processes that unfold over the course of recall, akin to the notion of how "compression," whereby recoding or chunking may result in less information needing to be retrieved, can boost memory during the recall phase by enhancing redintegration (Norris & Kalm, 2021) or reducing output interference induced by the chunked items on the remaining list items.

Although they did not test backward recall, Thalmann et al. (2019) found evidence that their pre-trained chunks facilitated memory during the study phase, by spatially cueing recall of one of two subsequence so they could counterbalance the order with which each subsequence was recalled. They found serial-position effects that were not changed when subsequence recall-order was varied, where if a chunk was studied earlier, later items were facilitated, but if the chunk was studied later, earlier items were not facilitated. There are several procedural differences between their methods and ours, any of which might explain why their effects were additionally found during study. Thalmann et al. (2019) used a chunk-training procedure whereby the first item of a chunk was only viewed and never recalled, whereas subsequent items were recalled; in our chunk-training, all items were recalled. They presented subsequences spatially segregated into rows, whether the subsequences contained a chunk or not and recall was spatially (and positionally) cued. In our method, lists were presented sequentially and centrally, with no positional cueing during study nor during recall. Their pre-trained chunks were recalled with near-ceiling accuracy apart from their third experiment whereas ours were well below ceiling on average. Finally, Thalmann et al. (2019) had no test of all-or-none retrieval, so it is possible that outside the all-or-none retrieval regime, pre-training of chunks may act more during the study phase.

Limitations of Transitional-Error Probabilities

Johnson (1969, 1970) makes a compelling argument for the idea that the presence of a spike in TEPs between recall of two subsequences is good evidence in support of all-or-none retrieval. However, as we alluded to in the introduction, the TEP measure is a bit strange in two chief ways. First, when one starts to work with TEPs, it immediately becomes evident that the TEP, itself, discards a lot of data. That is, whenever there is an error, the subsequent transition does not enter into the TEP calculation. If accuracy is low, it follows that there will be very little diagnostic data to decide whether or not retrieval of a subsequence (when it occurs) has the all-or-none characteristic. Second, at the opposite extreme, if recall is perfect, all TEP values will be zero, leaving no room to observe a spike between subsequences. Pragmatically, we were unable to devise an alternative measure that would make more use of the data. But all ways one could test for all-or-none retrieval would butt up against similar limitations, for conceptual reasons. At low levels of accuracy, subsequences or chunks are not sufficiently learnt to be retrieved all-or-none. At near-ceiling levels of accuracy, how is one to tell whether a perfectly recalled subsequence was recalled all-or-none, if it is almost always recalled correctly anyway? Ceiling and floor performance thus contribute very little diagnostic value to the question. Perhaps this intuition underlies Johnson's choice to compute TEPs collapsed across all degrees of learning in his multiple study/recall cycle design. Despite these limitations perhaps being unavoidable, the interpretation of the findings should be checked against these limitations. Accuracy was comfortably far from both floor and ceiling in both experiments in the aggregate (Figures S1 and S7). And in our TEP analyses, participants with missing values, which would have been due to ceiling accuracy, were excluded case-wise. In other words, the TEP functions we reported reflect data from participants at performance levels that have a dynamic range needed to compute the TEP function, itself. As always, there is a chance that excluded participants have some systematic characteristic in common beyond simply their performance level, so there may be unidentified boundary conditions on our findings.

Conclusion: subdivision of lists

Our knowledge *can* attain hierarchical-like organization (Bower, 1970), but still probably within a "flat" memory storage structure (see also arguments for non-hierarchical memory by Frank et al., 2012). If one learns labels as such, then access may be required to be sequential: access the chunk label "item," then access the chunk of component items referred to by that label. In the case of once-presented and once-recalled lists, with temporal grouping or overt instructions to subdivide, our findings, along with those of Liu and Caplan (2020), suggest that memory is quite non-hierarchical.

We suggest that conditions in which people subdivide lists invoke several very different mechanisms. Subdividing could show up as pauses between divisions (aka chunks, groups) or scalloped serial-position effects or might not show up at all. Subdividing a list does not inevitably make the memory task easier, but it can, especially at particular serial positions (Experiment 1 and Liu and Caplan, 2020), where the benefit can accumulate as recall proceeds. Sometimes chunks may be represented very much like isolated memories, and these could be hierarchically encoded or not. Sometimes sequences are recoded,

especially for stimuli such as digits and letters, which is orthogonal to any notion of hierarchical memory or even a chunk learned in isolation, but can produce memory retrieval that approximates all-or-none. Some subdivision effects may be due to explicit retrieval using within-chunk position, particularly when interpositions rates are high. Whereas subdivision of non-recodable word lists during the first exposure to a list produces recall that shows little evidence of all-or-none retrieval, recall resembles all-or-none when participants are exposed to a small amount of pre-training of constituent chunks, consistent with the idea that hierarchical recall is approximated by a non-hierarchical memory.

Declarations

This study was performed in line with the principles of the Declaration of Helsinki. Approval was granted by an Ethics Committee of the University of Alberta.

References

Anders, T. R., & Lillyquist, T. D. (1971). Retrieval time in forward and backward recall. *Psychonomic Science*, *22*(4), 205–206.

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, *38*, 341–380.

Anderson, J. R., & Matessa, M. (1997). A production system theory of serial memory. *Psychological Review*, *104*(4), 728–748.

Anderson, R. B., Crawford, J. C., & Bailey, M. H. (2022). Biasing the input: A yoked-scientist demonstration of the distorting effects of optional stopping on Bayesian inference. *Behavior Research Methods*, *54*(3), 1131–1147.

Barrouillet, P., Bernardin, S., & Camos, V. (2004). Time constraints and resource sharing in adults' working memory spans. *Journal of Experimental Psychology: General*, *133*(1), 83–100.

Bireta, T. J., Fry, S. E., Jalbert, A., Neath, I., & Surprenant, A. M. (2010). Backward recall and benchmark effects of working memory. *Memory & Cognition*, *38*(3), 279–291.

Bower, G. H. (1970). Organizational factors in memory. *Cognitive Psychology*, *1*, 18–46.

Bower, G. H., & Clark, M. C. (1969). Narrative stories as mediators for serial learning. *Psychonomic Science*, *14*(4), 181–182.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*(3), 539–576.

Brown, G. D. A., Preece, T., & Hulme, C. (2000). Oscillator-based memory for serial order. *Psychological Review*, *107*(1), 127–181.

Burgess, N., & Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review*, *106*(3), 551–581.

Chen, Z., & Cowan, N. (2005). Chunk limits and length limits in immediate recall: A reconciliation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1235–1249.

Cowan, N., Day, L., Saults, J. S., Keller, T. A., Johnson, T., & Flores, L. (1992). The role of verbal output time in the effects of word length on immediate memory. *Journal of Memory and Language*, *31*(1), 1–17.

Cowan, N., Saults, J. S., Elliott, E. M., & Moreno, M. V. (2002). Deconfounding serial recall. *Journal of Memory and Language*, *46*, 153–177.

Duncan, M., & Lewandowsky, S. (2005). The time course of response suppression: No evidence for a gradual release from inhibition. *Memory*, *13*(3/4), 236–246.

Ericsson, K. A., Chase, W. G., & Faloon, S. (1980). Acquisition of a memory skill. *Science*, *208*(4448), 1181–1182.

Ericsson, K. A., Delaney, P. F., Weaver, G., & Mahadevan, R. (2004). Uncovering the structure of a memorist's superior "basic" memory capacity. *Cognitive Psychology*, (3), 191–237.

Farrand, P., & Jones, D. (1996). Direction of report in spatial and verbal serial short-term memory. *Quarterly Journal of Experimental Psychology*, *49A*(1), 140–158.

Farrell, S. (2012). Temporal clustering and sequencing in short-term memory and episodic memory. *Psychological Review*, *119*(2), 223–271.

Farrell, S., & Lewandowsky, S. (2004). Modelling transposition latencies: Constraints for theories of serial order memory. *Journal of Memory and Language*, *51*(1), 115–135.

Frank, S. L., Bod, R., & Christiansen, M. H. (2012). How hierarchical is language use? *Proceedings of the Royal Society of London B*, *279*, 4522–4531.

Frankish, C. (1985). Modality-specific grouping effects in short-term memory. *Journal of Memory and Language*, *24*(2), 200–209.

Friendly, M., Franklin, P. E., Hoffman, D., & Rubin, D. C. (1982). The Toronto Word Pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods and Instrumentation*, *14*, 375–399.

Glanzer, M., & Fleishman, J. (1967). The effect of encoding training on perceptual recall. *Perception & Psychophysics*, *2*(12), 561–564.

Gobet, F., Lloyd-Kelly, M., & Lane, P. C. R. (2016). What's in a name? the multiple meanings of "chunk" and "chunking". *Frontiers in Psychology*, *7*(102).

Grenfell-Essam, R., & Ward, G. (2012). Examining the relationship between free recall and immediate serial recall: The role of list length, strategy use, and test expectancy. *Journal of Memory and Language*, *67*, 106–148.

Guérard, K., & Saint-Aubin, J. (2012). Assessing the effect of lexical variables in backward recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(2), 312–324.

Guérard, K., Saint-Aubin, J., Burns, S. C., & Chamberland, C. (2012). Revisiting backward recall and benchmark memory effects: A reply to Bireta et al. (2010). *Memory & Cognition*, *40*, 388–407.

Guitard, D., & Saint-Aubin, J. (2021). The irrelevant speech effect in backward recall is modulated by foreknowledge of recall direction and response modality. *Canadian Journal of Experimental Psychology*, *75*(3), 245–260.

Guitard, D., Saint-Aubin, J., Poirier, M., Miller, L. M., & Tolan, A. (2019). Forward and backward recall: Different visuospatial processes when you know what's coming. *Memory & Cognition*, *48*(5), 111–126.

Haberlandt, K., Thomas, J. G., Lawrence, H., & Krohn, T. (2005). Transposition asymmetry in immediate serial recall. *Memory*, *13*(3/4), 274–282.

Henson, R. N. A. (1998). Short-term memory for serial order: The Start-End Model. *Cognitive Psychology*, *36*(2), 73–137.

Henson, R. N. A. (1996, November). *Short-term memory for serial order* (Doctoral dissertation). University of Cambridge.

Hitch, G. J., Burgess, N., Towse, J. N., & Culpin, V. (1996). Temporal grouping effects in immediate recall: A working memory analysis. *Quarterly Journal of Experimental Psychology*, *49A*(1), 116–139.

Hulme, C., Roodenrys, S., Schweickert, R., Brown, G. D. A., Martin, S., & Stuart, G. (1997). Word-frequency effect on short-term memory tasks: Evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(5), 1217–1232.

JASP Team. (2023). JASP (Version 0.17)[Computer software]. https://jasp-stats.org

Johnson, N. F. (1969). Chunking: Associative chaining versus coding. *Journal of Verbal Learning and Verbal Behavior*, *8*(6), 725–731.

Johnson, N. F. (1970). The role of chunking and organization in the process of recall. In G. H. Bower (Ed.), *The psychology of learning and motivation advances in research and theory* (pp. 172–247). Academic Press.

Kahana, M. J., Mollison, M. V., & Addis, K. M. (2010). Positional cues in serial learning: The spin-list technique. *Memory & Cognition*, *38*(1), 92–101.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Society*, *90*(430), 773–795.

Kowialiewski, B., Gorin, S., & Majerus, S. (2021). Semantic knowledge constrains the processing of serial order information in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(12), 1958–1970.

Kowialiewski, B., Lemaire, B., & Portrat, S. (2020). How does semantic knowledge impact working memory maintenance? computational and behavioral investigations. *Journal of Memory and Language*, *117*(104208).

Kowialiewski, B., Lemaire, B., & Portrat, S. (2022). Between-item similarity frees up working memory resources through compression: A domain-general property. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001235

Lee, C. L., & Estes, W. K. (1981). Item and order information in short-term memory: Evidence for multilevel perturbation processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *7*(3), 149–169.

Li, S.-C., Chicherio, C., Nyberg, L., von Oertzen, T., Nagel, I. E., Papenberg, G., Sander, T., Heekeren, H. R., Lindenberger, U., & Bäckman, L. (2010). Ebbinghaus revisited: Influences of the BDNF Val66Met polymorphism on backward serial recall are modulated by human aging. *Journal of Cognitive Neuroscience*, *22*(10), 2164–2173.

Li, S.-C., & Lewandowsky, S. (1993). Intralist distractors and recall direction: Constraints on models of memory for serial order. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(4), 895–908.

Li, S.-C., & Lewandowsky, S. (1995). Forward and backward recall: Different retrieval processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 837–847.

Liu, Y. S., & Caplan, J. B. (2020). Temporal grouping and direction of serial recall. *Memory & Cognition*, *48*(7), 1295–1315.

Madigan, S. A. (1971). Modality and recall order interactions in short-term memory for serial order. *Journal of Experimental Psychology*, *87*(2), 294–296.

Manning, S. K. (1982). Forward and backward recall in the suffix paradigm. *Bulletin of the Psychonomic Society*, *20*(4), 199–202.

Manning, S. K., & Pacifici, C. (1983). The effects of a suffix-prefix on forward and backward serial recall. *American Journal of Psychology*, *96*(1), 127–134.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97.

Mızrak, E., & Oberauer, K. (2021). What is time good for in working memory? *Psychological Science*, *32*(8), 1325–1337.

Murdock, B. B. (1993). TODAM2: A model for the storage and retrieval of item, associative, and serial-order information. *Psychological Review*, *100*(2), 183–203.

Murdock, B. B. (1995). Developing TODAM: three models for serial-order information. *Memory & Cognition*, *23*(5), 631–645.

Murdock, B. B. (2005). Storage and retrieval of serial-order information. *Memory*, *13*(3/4), 259–266.

Norris, D., & Kalm, K. (2021). Chunking and data compression in verbal short-term memory. *Cognition*, *208*(104534).

Oberauer, K. (2003). Understanding serial position curves in short-term recognition and recall. *Journal of Memory and Language*, *49*, 469–483.

Oberauer, K. (2022). When does working memory get better with longer time? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(12), 1754–1774.

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, R., M. R.and Höchenberger, Sogo, H., Kastman, E., & Lindeløv, J. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *5*(1), 195–203.

Popov, V., & Reder, L. (2020). Frequency effects on memory: A resource-limited theory. *Psychological Review*, *127*(1), 1–46.

Ritchie, G., Tolan, G. A., Tehan, G., Goh, H. E., Guérard, K., & Saint-Aubin, J. (2015). Phonological effects in forward and backward serial recall: Qualitative and quantitative differences. *Canadian Journal of Experimental Psychology*, *69*(1), 95–103.

Ryan, J. (1969a). Grouping and short-term memory: Different means and patterns of grouping. *Quarterly Journal of Experimental Psychology*, *21*(2), 137–147.

Ryan, J. (1969b). Temporal grouping, rehearsal and short-term memory. *Quarterly Journal of Experimental Psychology*, *21*(2), 148–155.

Spurgeon, J., Ward, G., Matthews, W. J., & Farrell, S. (2015). Can the effects of temporal grouping explain the similarities and differences between free recall and serial recall? *Memory & Cognition*, *43*(3), 469–488.

St. Clair-Thompson, H. L., & Allen, R. J. (2013). Are forward and backward recall the same? a dual-task study of digit recall. *Memory & Cognition*, *41*, 519–532.

Tan, L., Ward, G., Paulauskaite, L., & Markou, M. (2016). Beginning at the beginning: Recall order and the number of words to be recalled. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(8), 1282–1292.

Thalmann, M., Souza, A. S., & Oberauer, K. (2019). How does chunking help working memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(1), 37–55.

Thomas, J. G., Milner, H. R., & Haberlandt, K. F. (2003). Forward and backward recall: Different response time patterns, same retrieval order. *Psychological Science*, *14*(2), 169–174.

Ward, G., & Tan, L. (2019). Control processes in short-term storage: Retrieval strategies in immediate recall depend upon the number of words to be recalled. *Memory & Cognition*, *47*(4), 658–682.

Ward, G., & Tan, L. (2023). The role of rehearsal and reminding in the recall of categorized word lists. *Cognitive Psychology*, *143*(101563).

Watkins, M. J., & Bloom, L. C. (1999). *Selectivity in memory: An exploration of willful control over the remembering process* [Unpublished manuscript].

Wickelgren, W. A. (1967). Rehearsal grouping and hierarchical organization of serial position cues in short-term memory. *Quarterly Journal of Experimental Psychology*, *19*(2), 97–102.

Worthen, J. B., & Hunt, R. R. (2008). Mnemonics: Underlying processes and practical applications (J. H. Byrne, Ed.). *Learning and memory: A comprehensive reference*, *2*, 145–153.

Effect	F	df, error df	MSE	p	η_p^2	BF
Condition	2.205	2, 265	5.741	0.112	0.016	0.132
Direction	0.002	2, 265	5.741	0.967	<0.001	0.097
Condition×Direction	1.062	2, 265	5.741	0.347	0.008	0.014

Table S1

Experiment 1: ANOVA on mean initiation time with design Condition[Control,Subjective Chunking, Temporal Grouping]×Direction[Forward,Backward]. $BF = BF_{inclusion}$.

Supplementary Materials

Experiment 1: Serial-position curves

Figure S1 plots the serial-position curves for Experiment 1 (top) alongside those for the first experiment of Liu and Caplan (2020).

Figure S2 plots inter-response times (and initiation times, reported in Table S1), exhibiting very little evidence of a pause between chunks in the forward direction, but contains evidence of such a pause in the backward direction for the Subjective Chunking and Temporal Grouping conditions but not Control ($p<0.005$).

Experiment 2: Baseline serial recall phase

For the Baseline phase (serial recall of 6-word lists alternating forward/backward/forward/backward), accuracy is plotted in Figure S3, both mean and median (the very small number of data points going into each average makes the mean highly sensitive to outliers; median is more stable) inter-response times in Figure S4 and transitional-error probabilities in Figure S5. Accuracy shows the typical primacy-dominance in the forward direction, which reverses in the backward direction, suggesting that output order is the major determinant of accuracy. Inter-response times are also primarily driven by output order rather than serial position. Very little trace of scalloping is evident in the accuracy serial-position curves, and there is no clear increase in inter-response time that might indicate a between-group pause that might be consistent across participants. Similarly, although the transitional-error probabilities differ between the forward and backward direction, no characteristic TEP spike pattern is visible, so the participants in Experiment 2 did not, on the whole, start the session with all-or-none retrieval or chunking evident in their behaviour.

Experiment 2: Chunk training phase

Figure S6 plots the learning curves (accuracy as a function of training cycle) for the 4×3 and 1×12 sets. Both sets show evidence of learning over the three cycles. For the 4×3 set, despite the inclusion of an end-of-list distractor task, accuracy is near-ceiling, presumably due to the very short list length. The degree of learning may thus be overestimated for

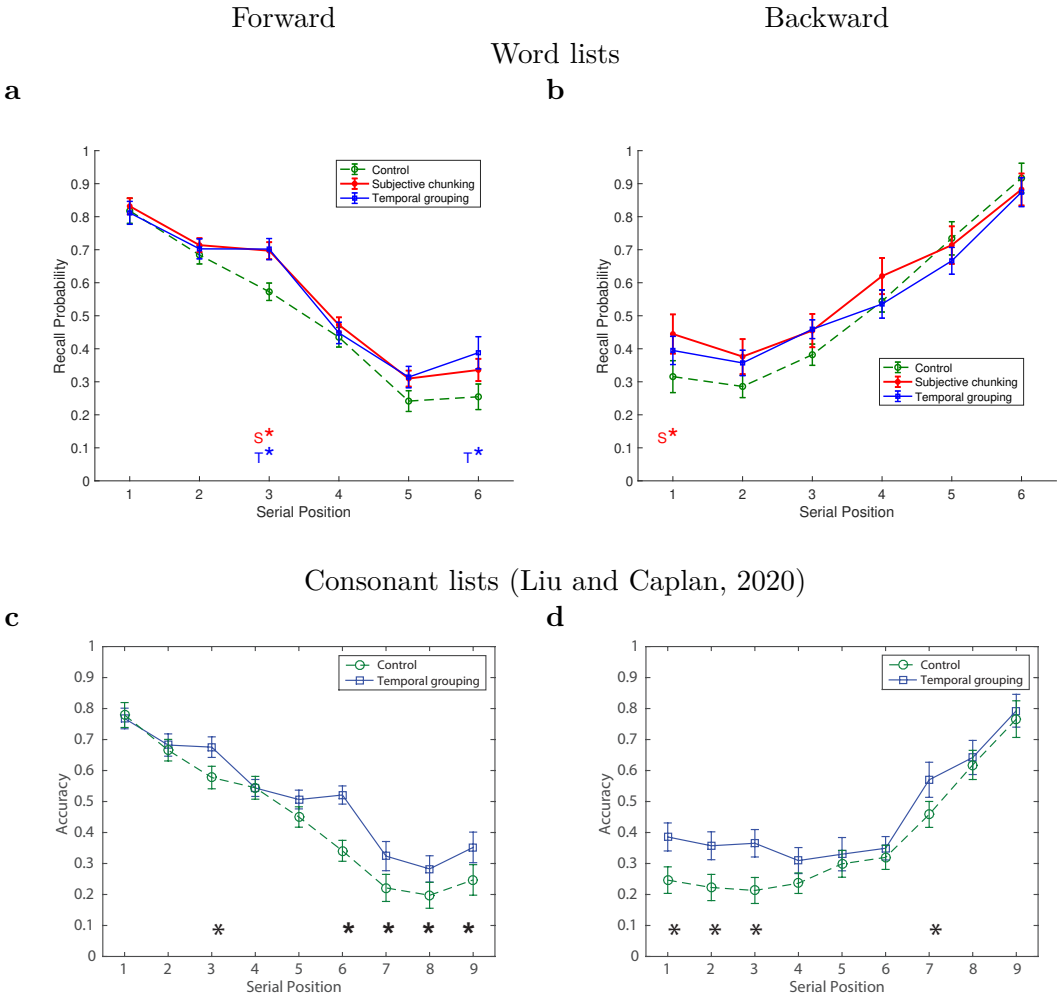

Figure S1

Experiment 1: Accuracy as a function of serial position for forward (a) and backward (b) recall directions. For comparison, the corresponding serial-position curves are adapted from Liu and Caplan (2020), Experiment 1 (recall direction between-subjects) in panels c and d, respectively. Error bars plot 95% confidence intervals based on standard error of the mean. In (c) and (d), asterisks denote significant post-hoc t tests between grouped and control lists.

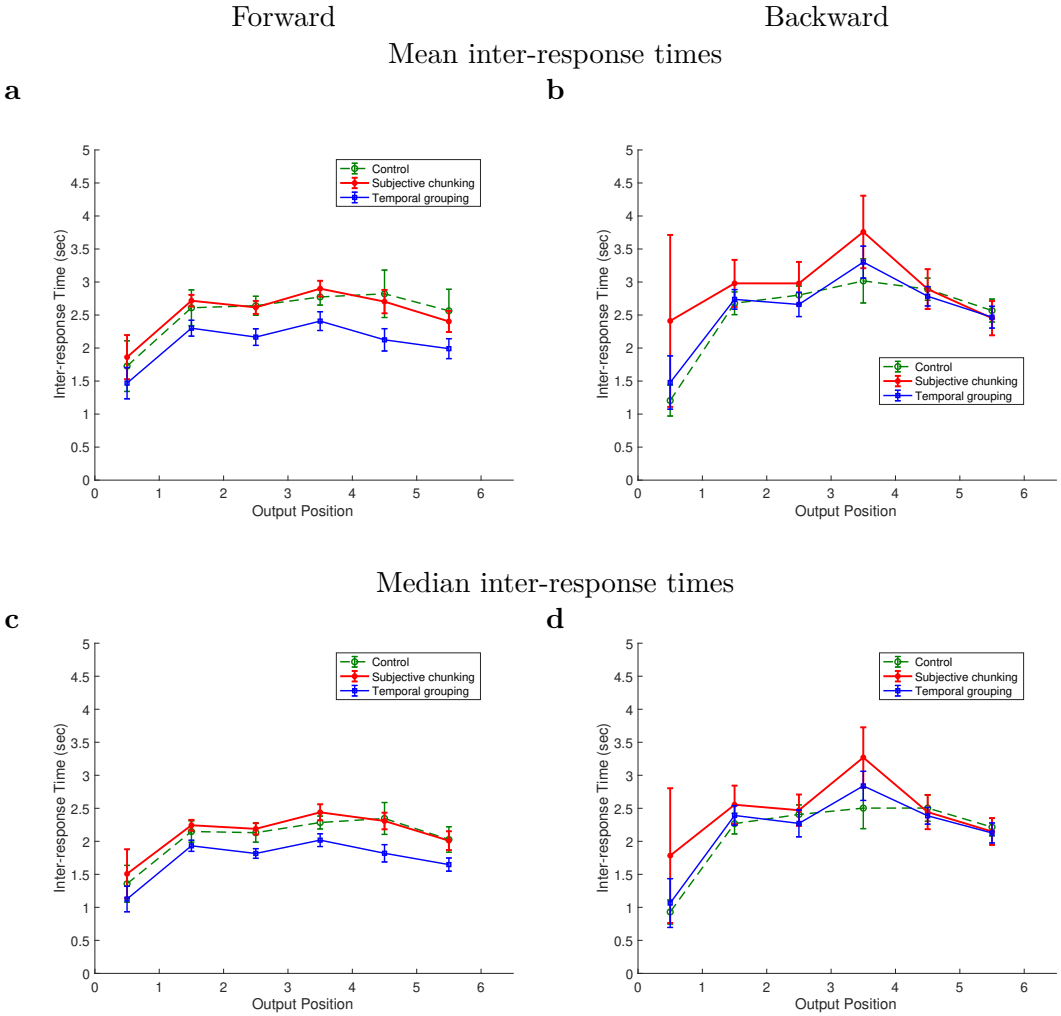

Figure S2

Experiment 1: Inter-response time (onset-to-onset) for correct recalls as functions of output position, for forward (a) and backward (b) recall directions. Points are plotted between the two output positions comprising the transition. Error bars plot 95% confidence intervals based on standard error of the mean. Initiation times are often longer, particularly in forward recall; curiously, initiation times in this experiment were shorter than inter-response times. ***The first point plots initiation time.***

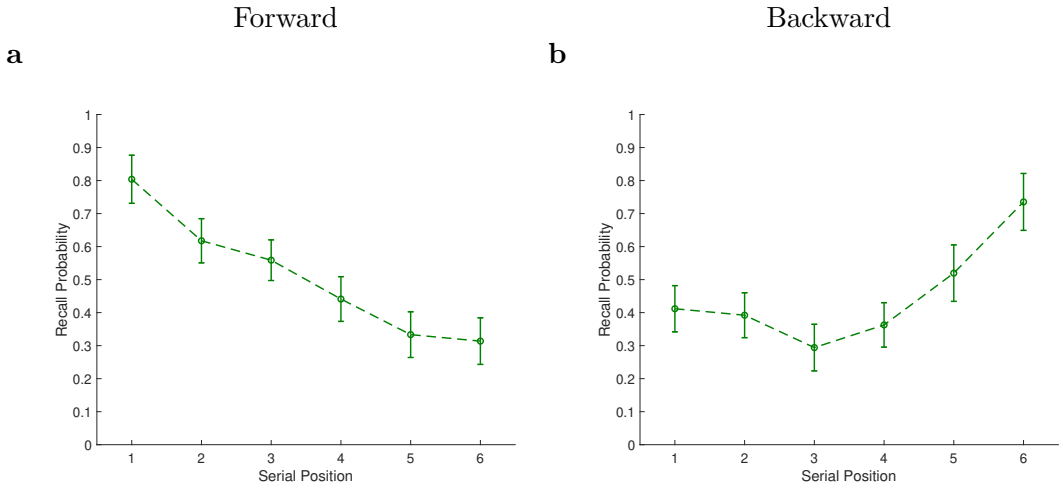

Figure S3

Experiment 2: Baseline-list recall accuracy as a function of serial position for forward (a) and backward (b) recall directions. Error bars plot 95% confidence intervals based on standard error of the mean.

Effect	F	df, error df	MSE	p	η_p^2	BF
Training set	<0.001	1, 50	3.359	0.984	<0.001	0.134
Direction	1.655	1, 50	6.281	0.204	0.032	0.332
Training set×Direction	0.127	1, 50	2.799	0.723	0.003	0.044

Table S2

Experiment 2: ANOVAs on mean initiation time with design Training set[4×3, 1×12]×Direction[Forward, Backward]. BF = $BF_{inclusion}$.

the 4×3 condition; in the probe lists derived from the trained chunks, the 4×3 and 1×12 chunks were nearly equally well remembered (Figure S7 and S8).

Experiment 2: Probe-List phase

Figure S7 plots the accuracy serial-position curves for the probe-list phase of Experiment 2. Serial-position effects were quite similar when derived from the 1×12 set as from the 4×3 set, and approximately mirror-image effects for backward compared to forward recall, suggesting that accuracy was largely driven by output effects rather than serial-position effects.

Figure S8 plots inter-response time serial-position curves, which exhibit evidence of pausing between chunks with mean (top) measures across both recall directions, except for the last chunk in forward direction ($p<0.05$). The two source sets are largely similar and backward serial-position effects highly resemble forward. Analysis of initiation times is reported in Table S2

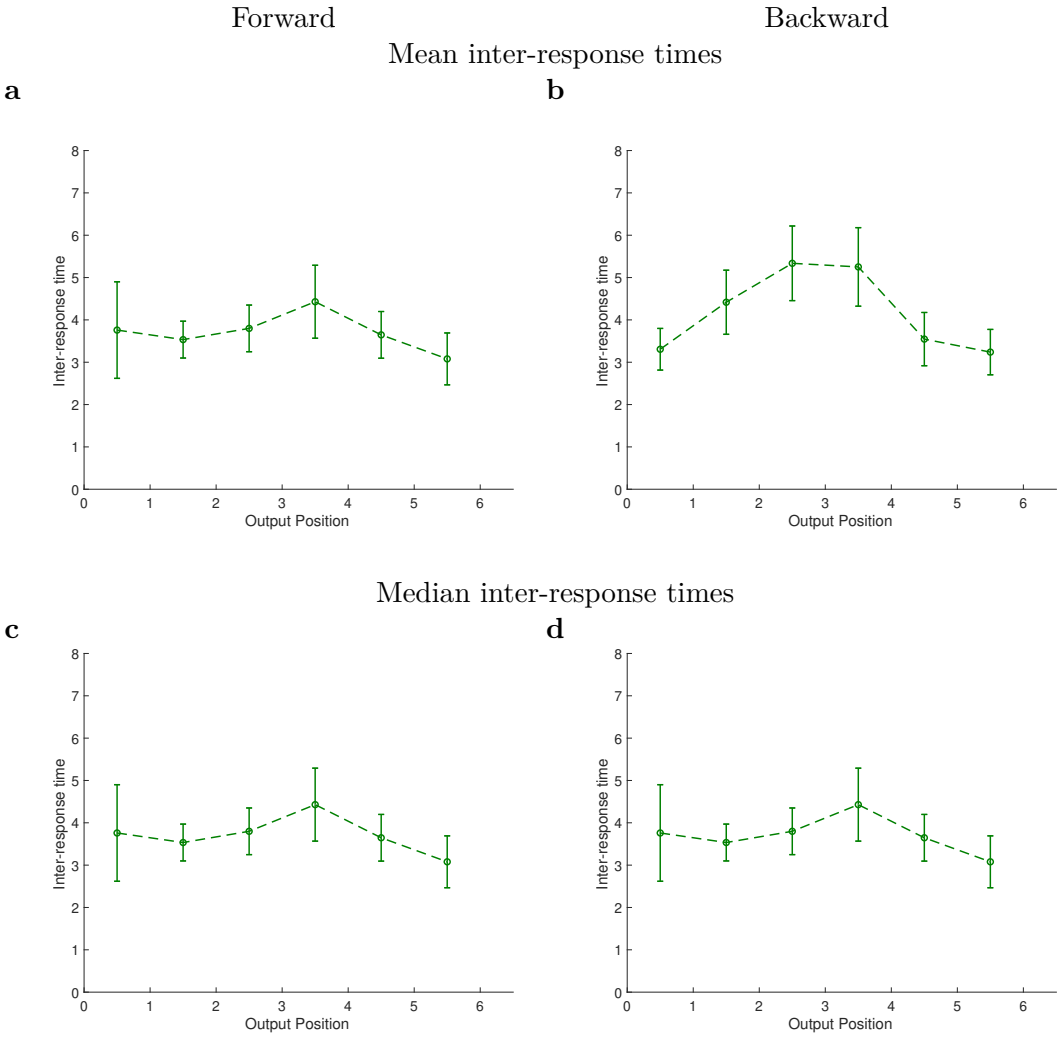

Figure S4

Experiment 2: Baseline-list recall mean (row 1) and median (row 2) inter-response times for all responses as a function of output transitions for forward (a,c) and backward (b,d) recall directions. Error bars plot 95% confidence intervals based on standard error of the mean.

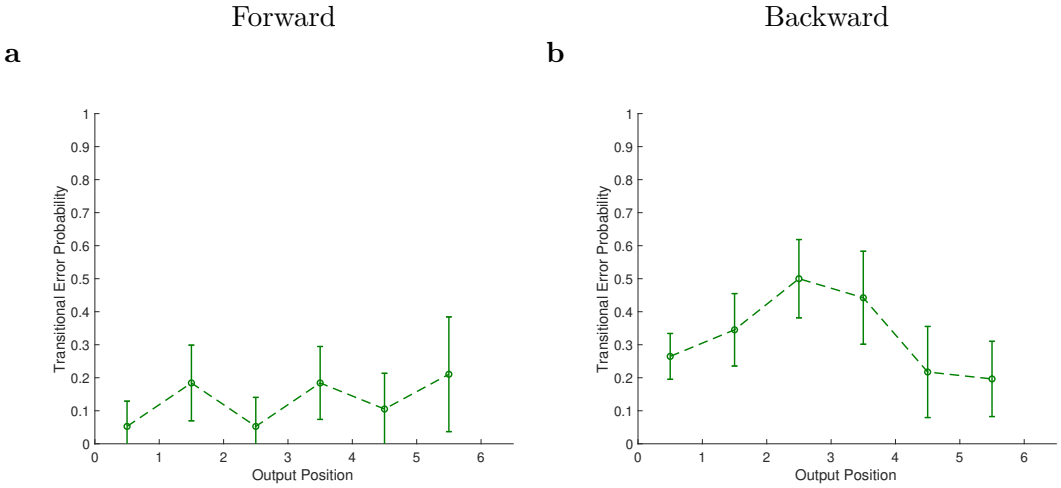

Figure S5

Experiment 2: Baseline-list Transitional-Error Probabilities in Experiment 2 as a function of transition for forward (a) and backward (b) recall directions. Participants with one or more missing values were omitted. Error bars plot 95% confidence intervals based on standard error of the mean. ***The first point plots initiation error probability.***

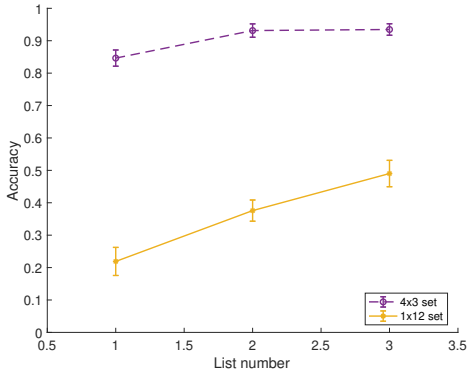

Figure S6

Experiment 2: Training phase serial recall accuracy (proportion correct) as a function of cycle number for each training set. Error bars plot 95% confidence intervals based on standard error of the mean.

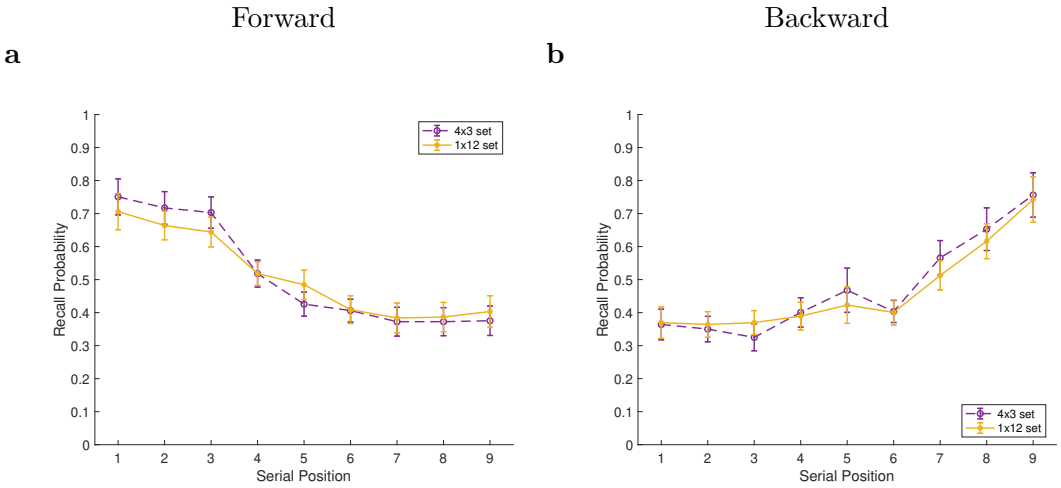

Figure S7

Experiment 2: Probe-list recall accuracy as a function of serial position for forward (a) and backward (b) recall directions. Error bars plot 95% confidence intervals based on standard error of the mean.

Following up on Figure 6, Figure S9 reports the TEP analyses with all participants included (apart from those excluded for ceiling or floor performance), excluding missing values cell-wise. For both training sets, TEPs consistent with all-or-none retrieval can be seen within the first and last recalled chunks in both directions, but not the middle chunk. However, it is important consider that the middle chunk, especially, may be obscured by different subsets of participants contributing to each point.

For this reason, we find most informative Figure 6 included in the main text, where participants were removed entirely if they had one or more missing values. Because this removes a considerable number of participants, we also report in Figure S10 an intermediate approach, where each chunk was analyzed separately and participants were excluded from the analysis of a given chunk if they had missing values for that chunk, disregarding the other chunks. These should still be interpreted with some caution, as different subsets of participants contribute to the plots of each chunk (but do not vary within-chunk).

Omitted-chunk examples

On 10 lists across 6 of the participants, all three items of the omitted chunk of the source set were recalled together. On all 10 out of 10 of these recalls, the omitted chunk was recalled in correct order of report (forward when recall was forward, or backward when recall was backward), and aligned to a natural chunk boundary. Table S3 reports each of these.

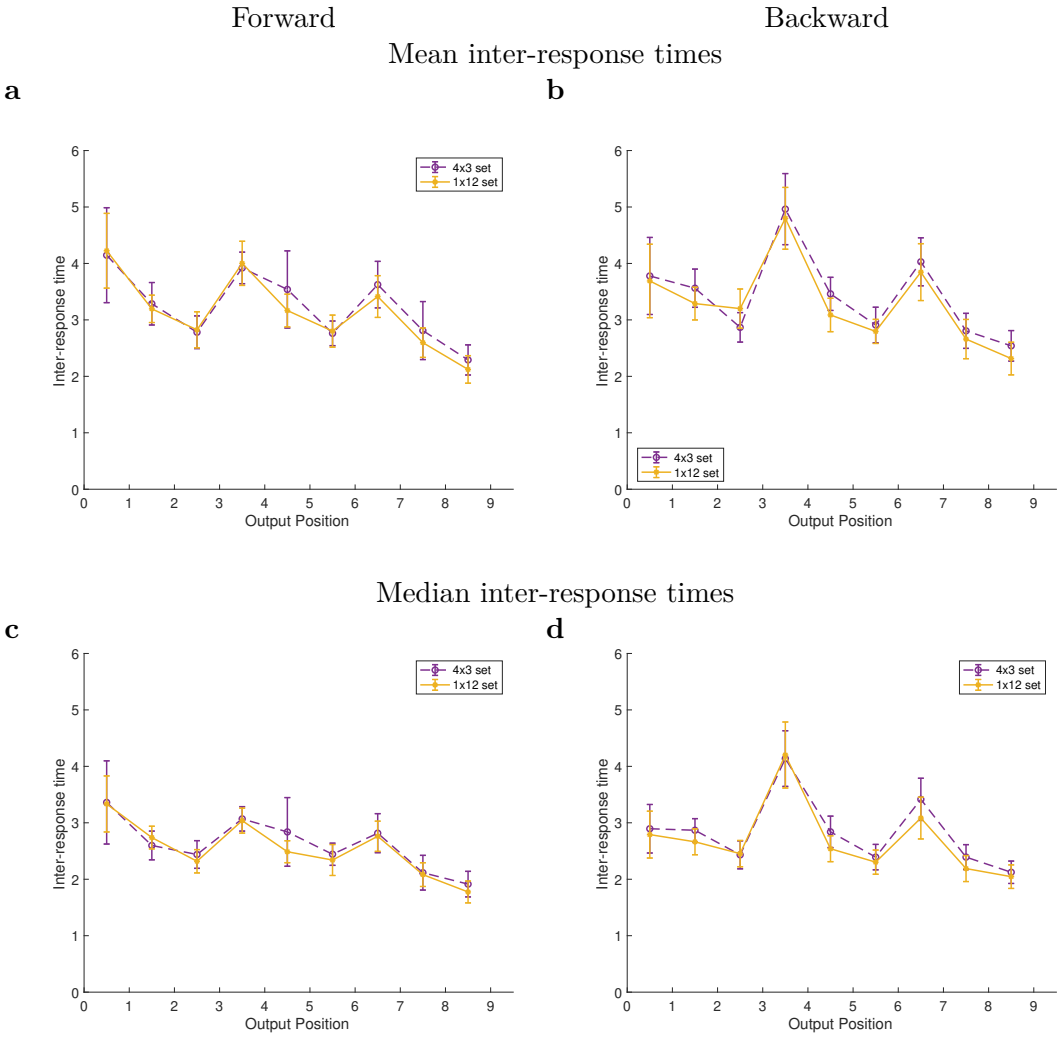

Figure S8

Experiment 2: Mean (row 1) and median (row 2) inter-response times for all responses as a function of output transitions for forward (a,c) and backward (b,d) recall directions. Error bars plot 95% confidence intervals based on standard error of the mean. ***The first point plots initiation time.***

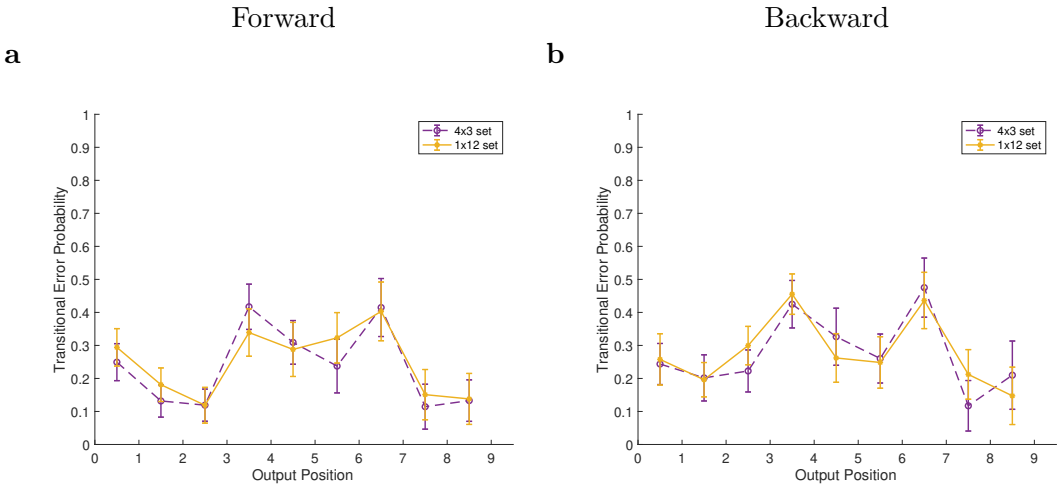

Figure S9

Experiment 2: Transitional-Error Probabilities as a function of transition for forward (a) and backward (b) recall directions. Error bars plot 95% confidence intervals based on standard error of the mean. Unlike Figure 6, missing values were excluded cell-wise rather than removing the entire participant. ***The first point plots initiation error probability.***

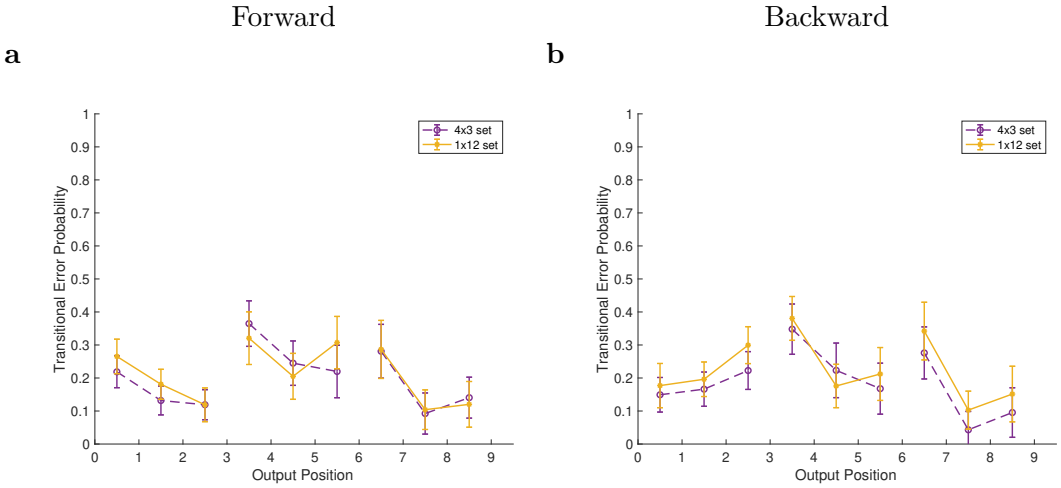

Figure S10

Experiment 2: Transitional-Error Probabilities as a function of transition for forward (a; N=49/43/35 and N=49/43/33 for the 4×3 and 1×12 sets, chunks 1/2/3, respectively) and backward (b; N=45/36/29 and N=46/37/36 across 4×3 and 1×12 sets, chunk 1-2-3, respectively) recall directions. Error bars plot 95% confidence intervals based on standard error of the mean. ***The first point plots initiation error probability.***

Forward, 1×12									
Omitted	KEEPER	VALUE	DANGER						
Studied:	DIET	MEETING	ONION	ACRE	SPIDER	MISTRESS	LEADER	MAKER	TUNNEL
Recalled:	**KEEPER**	**VALUE**	**DANGER**	DIET	MEETING	ONION	LEADER	MAKER	TUNNEL
Backward, 1×12									
Omitted	ACRE	SPIDER	MISTRESS						
Studied:	KEEPER	VALUE	DANGER	LEADER	MAKER	TUNNEL	DIET	MEETING	ONION
Recalled:	TUNNEL	MAKER	LEADER	**MISTRESS**	**SPIDER**	**ACRE**	DANGER	VALUE	KEEPER
Backward, 4×3									
Omitted	OBJECT	JUSTICE	MISCHIEF						
Studied:	CLOSET	BUBBLE	THUNDER	TURKEY	ANGLE	BARGAIN	SPEAKER	SENATE	COUPLE
Recalled:	**MISCHIEF**	**JUSTICE**	**OBJECT**	BARGAIN	ANGLE	THUNDER	COUPLE	BUBLE	CLOSET
Forward, 4×3									
Omitted	IRON	POWDER	MARBLE						
Studied:	PLAYER	FATHER	MILLION	BASIS	NOVEL	THEORY	MEANING	COTTON	MANNER
Recalled:	PLAYER	FATHER	MILLION	BASIS	NOVEL	THEORY	**IRON**	**POWDER**	**MARBLE**
Forward, 4×3									
Omitted	PRINCESS	SENATE	MAKER						
Studied:	BULLET	VALUE	MISTAKE	VELVET	ACCOUNT	SENTENCE	CLUSTER	AMOUNT	TRAITOR
Recalled:	VELVET	ACCOUNT	SENTENCE	**PRINCESS**	**SENATE**	**MAKER**	PASS	PASS	PASS
Backward, 1×12									
Omitted	METHOD	NEEDLE	COUSIN						
Studied:	MEADOW	LAWYER	BELIEF	HUNTER	INCOME	RIDER	OYSTER	FEATURE	JOURNAL
Recalled:	JOURNAL	FEATURE	OYSTER	RIDER	INCOME	HUNTER	**COUSIN**	**NEEDLE**	**METHOD**
Forward, 4×3									
Omitted	SULPHUR	MISTRESS	HATRED						
Studied:	TUNNEL	FINGER	SAILOR	EAGLE	RESOURCE	WOMEN	KINDNESS	ONION	PALACE
Recalled:	**SULPHUR**	**MISTRESS**	**HATRED**	EAGLE	RESOURCE	PLAYER	TUNNEL	FINGER	SAILOR
Backward, 4×3									
Omitted	FORTUNE	NATION	LAYER						
Studied:	SAILOR	COMPOUND	HEAVEN	RECEIPT	MONSTER	SANDWICH	LOVER	PROVINCE	COUNTY
Recalled:	COUNTY	PROVINCE	LOVER	SANDWICH	MONSTER	RECEIPT	**LAYER**	**NATION**	**FORTUNE**
Forward, 4×3									
Omitted	LOVER	PROVINCE	COUNTY						
Studied:	RECEIPT	MONSTER	SANDWICH	FORTUNE	NATION	LAYER	SAILOR	COMPOUND	HEAVEN
Recalled:	RECEIPT	MONSTER	SANDWICH	FORTUNE	NATION	LAYER	**LOVER**	**PROVINCE**	**COUNTY**
Forward, 1×12									
Omitted	ACCOUNT	SCHOLAR	MEADOW						
Studied:	TIGER	CHAPEL	RECORD	MESSAGE	BANNER	ARTIST	DAUGHTER	TIMBER	HERALD
Recalled:	**ACCOUNT**	**SCHOLAR**	**MEADOW**	PASS	PASS	PASS	DAUGHTER	TIMBER	HERALD

Table S3

Experiment 2: The 10 examples of all words of the omitted chunk being recalled on a single list. Forward/Backward denotes recall direction. 1×12 or 4×3 denotes source-list condition. Horizontal lines separate distinct participants; examples with no line separation were produced by the same participant. In these examples, all recalls apart from the omitted chunk were correct-in-position unless underlined (including PASS and BUBLE, an item error, albeit minor). In the recall sequence, omitted-chunk words are denoted in boldface.

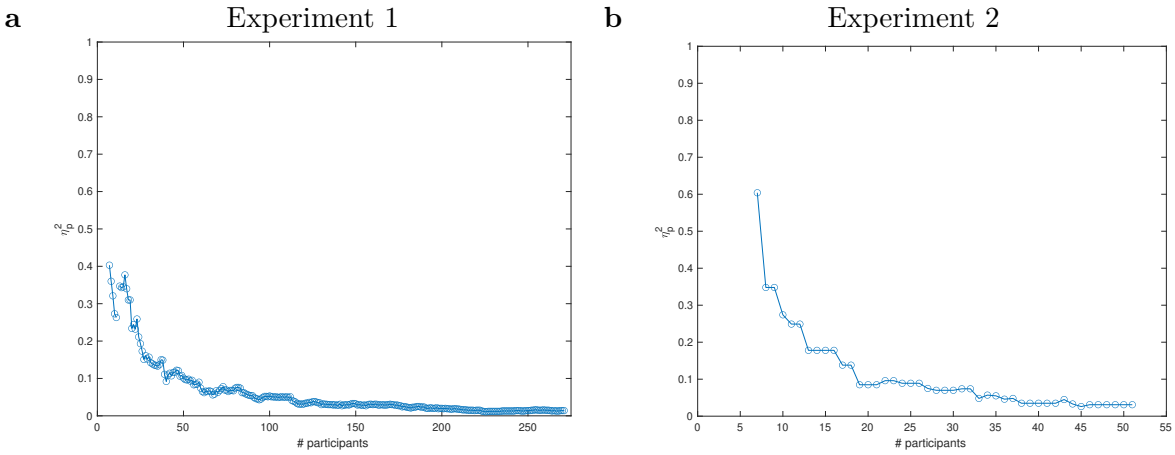

Figure S11

Stability analysis for Experiment 1 (a) and Experiment 2. Following R. B. Anderson et al. (2022), the effect size, η_p^2, is plotted as a function of number of participants, as participants are added one at a time, to check whether the effect size has stabilized by the time data-collection stopped. For Experiment 1, we plot the effect size for the three-way interaction, Output Position×Condition×Direction of strict-scored accuracy which ended with a supported null Bayes Factor (under 1:3). For Experiment 2, we plot the effect size for the three-way interaction Transition×Training set×Direction on Transitional-Error Probabilities), which also had supported null Bayes Factor (under 1:3).