Article:

Allen, Richard J., Hitch, Graham James and Baddeley, Alan David orcid.org/0000-0002-9163-0643 (2018) Exploring the sentence advantage in working memory:Insights from serial recall and recognition. Quarterly Journal of Experimental Psychology. pp. 1-15. ISSN 1747-0226

https://doi.org/10.1177/1747021817746929

Exploring the sentence advantage in working memory: Insights from serial recall and recognition

Richard J. Allen[1], Graham J. Hitch[2] & Alan D. Baddeley[2]

1. School of Psychology, University of Leeds, UK
2. Department of Psychology, University of York, UK

IN PRESS, QJEP

Correspondence concerning this article should be addressed to Richard Allen, School of Psychology, University of Leeds, LS2 9JT, United Kingdom. E-mail: r.allen@leeds.ac.uk

Abstract

Immediate serial recall of sentences has been shown to be superior to that of unrelated words. The present study was designed to further explore how this effect might emerge in recall, and to establish whether it also extends to serial recognition, a different form of response task that has relatively reduced output requirements. Using auditory or visual presentation of sequences, we found a substantial advantage for sentences over lists in serial recall, an effect shown on measures of recall accuracy, order, intrusion, and omission errors and reflected in transposition gradients. In contrast however, recognition memory based on a standard change detection paradigm gave only weak and inconsistent evidence for a sentence superiority effect. However, when a more sensitive staircase procedure imported from psychophysics was used, a clear sentence advantage was found although the effect sizes were smaller than those observed in serial recall. These findings suggest that sentence recall benefits from automatic processes that utilise long-term knowledge across encoding, storage, and retrieval.

Keywords: working memory; sentence memory; recall; recognition; staircase method

Exploring the sentence advantage in working memory: Insights from serial recall and recognition

Since its introduction by Jacobs (1887), immediate serial recall has continued to feature, both as a measure of individual cognitive differences as initially proposed, and because of its theoretical relevance. This stems from its potential to throw light on two fundamental questions of human memory, namely how serial order is stored, together with the question of the mechanisms underlying chunking, the capacity to increase span by combining items into larger stored units.

These two questions have tended to generate separate literatures. Studies focused on explanations of maintaining serial order, have tended to use a limited pool of isolated items, digits, letters or individual words (see Hurlstone, Hitch & Baddeley, 2014, for a review). When words are used, it is their phonological characteristics that dominate recall, in contrast to long-term serial learning where semantic features are more important (Baddeley, 1966a, 1966b). The role of chunking is most clearly seen using sentences, typically tested using an open set in which sentences are not repeated. This can lead to a substantial increase in span, for example from five unrelated words to around 15 words within a sentence (Brener, 1940; Baddeley, Allen & Hitch, 2009). Sentence recall has been extensively explored by Potter and colleagues (Lombardi & Potter, 1992; Potter & Lombardi, 1990, 1998) who propose a “regeneration of the sentence from a conceptual representation using words that have recently been activated” (Potter & Lombardi, 1998, p.633). They (Potter & Lombardi, 1990) observed that potentially disruptive synonyms presented along with the sentence intrude only when they are consistent with the underlying theme of the sentence, proposing that recall will be verbatim only when the right set of words is activated. Following this, effects of syntactic (Potter & Lombardi, 1998) and phonological priming (Schweppe, Rummer, Bormann, & Martin, 2011; see also Rummer & Engelkamp, 2001, 2003) at retrieval have also been observed in certain contexts, while Schweppe and Rummer (2007) suggested that morphosyntactic representations may also contribute. Thus, within this approach, a sentence is primarily represented via its meaning or gist, extracted during comprehension, and this propositional representation drives subsequent regeneration of the sequence at recall via priming at lexico-semantic and phonological levels. List memory, in contrast, depends on the storage and preservation of a surface phonological representation.

Stored language knowledge has also been suggested to impact on immediate memory tasks via a process of redintegration. This term has been proposed to refer to the influence of existing knowledge on the capacity to recall a word or letter sequence (Schweickert, 1993). There are however two possible modes within which such knowledge can be used. One form of redintegration is more akin to pattern completion and is likely to operate in a bottom-up manner at the lexical level on recall of individual items. The other may represent an explicit, top-down, conceptually-driven strategy to utilise such knowledge, and may be closer in nature to the conceptually-based regeneration proposed by Potter and Lombardi (1998). However, unlike Potter's approach, which assumes different types of representation underlying sentences and lists, processes of redintegration may primarily operate during recall. Under the retrieval-based redintegration hypothesis (Saint-Aubin & Poirier, 2000), for example, items are held in short-term memory until recall, at which point they are repaired by comparison with long-term knowledge. This kind of process, applied in a top-down manner via conceptual and syntactic knowledge, might provide an additional account of sentence superiority effects in immediate serial recall.

One problem in comparing span for unrelated items with sentence span concerns the substantial difference between them in length, typically five words versus around 15. To study the influence of a concurrent attentional load on memory for lists and sentences, Baddeley et al. (2009; Allen & Baddeley, 2008) devised what they termed *constrained sentences*. These were a halfway house between lists and sentences in that they repeatedly used a limited set of words but combined them into meaningful sentences, a process that involved both consistent syntactic structure and a conceptual structure that was impoverished. For example, one sentence might be, *tall soldier follows waiter and old sad teacher*, followed by, *teacher or tall sad bishop meets fat waiter.* Under these conditions, span was reduced to about two words beyond span for the equivalent words in random order.

Baddeley et al. (2009) compared immediate recall of lists and sentences under a range of conditions selected to limit the use of the various proposed components of working memory. We proposed that if the sentence advantage was attributable to attentionally driven forms of chunking or redintegration, then the sentence advantage should be reduced. Recall for both sequence types exhibited a small effect of concurrent visual processing, a larger effect of articulatory suppression, preventing verbal rehearsal, with the largest effect occurring when rehearsal was prevented by an attentionally demanding task. However, there was no suggestion of an interaction between concurrent task and sequence type. These patterns emerged across auditory and visual presentation modalities, indicating that, for recall at least, the nature of the perceptual input does not mediate the sentence effect. While these findings are also broadly consistent with theoretical approaches such as the conceptual regeneration hypothesis (e.g. Potter & Lombardi, 1990), we concluded that the sentence advantage was an automatic feature of the involvement of long-term language structure, resulting in more effective binding of information within the episodic buffer (Baddeley, 2000; Baddeley, Allen, & Hitch, 2011).

The experiments that follow focus on establishing the generality of previously observed sentence superiority effects. A primary question is whether the sentence advantage operates at the level of retrieval, or at some other stage of processing. This was prompted by several studies concerned with the effects of lexicality on immediate serial recall and recognition. When the item retrieval process is minimised by using recognition memory paradigms (in which the sequence is re-presented and participants decide whether a change has occurred), lexicality effects appear to be reduced or absent (Gathercole. Pickering, Hall, & Peaker, 2001, Hulme, Roodenrys, Schweickert, Brown, Martin, & Stuart, 1997; Jefferies et al., 2006; Macken, Taylor, & Jones, 2014). The present experimental series therefore examines whether the sentence effect is also diminished or removed when moving to a recognition task that places relatively reduced demands on retrieval processes. In addition, there is evidence that the presence or absence of a lexicality effect in recognition may depend on whether presentation is auditory or visual (Macken et al., 2014). Effects appear to be attenuated following auditory presentation, with Macken et al. attributing this to the use of speech-based pattern-matching during the recognition phase. We therefore use our constrained sentences and lists to compare performance on immediate recall and recognition using both auditory (Experiments 1-2) and visual (Experiments 3-4) presentation.

Experiment 1

Memory for aurally presented short sequences of words that differed only in the presence or absence of a meaningful structure was assessed, using immediate serial recall and recognition of constrained sentences and word lists. Constrained sentences provide a tightly controlled comparison with word lists as both sets of materials are derived from the same small pool of nouns, verbs, adjectives and function words. In constrained sentences a subset of items from the pool are presented in a grammatical order whereas in word lists, items selected using the same constraints are presented in an ungrammatical order. Constrained sentences are meaningful, despite being somewhat unnatural, and lead to a highly reliable sentence superiority effect in immediate serial recall (Baddeley et al. (2009). Using such well-controlled materials allowed us to compare detailed aspects of serial recall such as serial position curves and transposition gradients (e.g. Hurlstone et al., 2014; Nairne, 1992). These details of performance are of interest because an episodic buffer account generates an expectation of similarities between sentences and word lists given a common binding and the suggestion that one of its functions concerns serial ordering (Burgess & Hitch, 2005). If on the other hand sentence and list recall depend on fundamentally different representations, as envisaged by the conceptual regeneration hypothesis, we might expect more substantial differences in the details of serial recall between these sequence types.

To assess serial recognition, we used a version of the task in which change trials featured switches in word position within sequences that maintained the meaningful grammatical structure of sentences (i.e. noun switched with noun, adjective with adjective, and verb with function word). In the recognition conditions, each presentation was immediately followed by an identical or changed sequence, with participants simply asked to detect whether a change in order had occurred. This was assumed to reduce the explicit retrieval demands placed by the more complex response phase required by immediate serial recall. Our assumption was that if the sentence advantage emerges through processes operating during encoding and storage, substantial effects over list recall should be observed in both recall and recognition measures. However, if we adopt a strict retrieval-based account and assume that this effect is solely due to processes (such as redintegration) operating exclusively during overt recall, then little or no impact of sequence type should be observed in recognition.

Method

Participants

Twenty undergraduate and postgraduate students (12 female) from the University of York took part, in return for course credit or financial payment. All participants in this and subsequent experiments were native English speakers.

Materials

Testing was controlled on a Macintosh computer using a SuperCard program, with stimuli presented through headphones and a 15” monitor. A single set of 15 words, consisting of four nouns (bishop, soldier, teacher, waiter), four adjectives (fat, old, sad, tall), four verbs (follows, helps, insults, meets), and three function words (and, not, or) was used in each of the conditions, in all reported experiments. From this set, two sets of 2 practice sequences and 12 test sequences were each constructed for the sentence and list conditions. Each sequence consisted of 8 words (3 nouns, 3 adjectives, 1 verb, 1 function word). The order of words within sentences was grammatically correct (e.g. *FAT WAITER HELPS TALL SOLDIER NOT OLD BISHOP*), whereas the order of words within lists was always at least two positional switches removed from being correct (e.g. *TALL HELPS NOT WAITER BISHOP FAT OLD SOLDIER*). The absolute positional constraints that arose in the sentence sequences were also imposed on list sequences, in that the first word in a sequence was always a noun or adjective, the penultimate word was never a noun, and the final word was always a noun.

Each of the words in the experiment were individually recorded and digitized by a male English speaker. The same individual audio files were used in all conditions, meaning that sequences contained no speech cues such as prosody or co-articulation.

Design and Procedure

Each participant performed all four conditions in turn (serial recall vs. serial recognition; sentence vs. list), with order counter-balanced across the experiment. The two sets of sentences and lists were presented equally often in recall and recognition conditions. Participants were informed at the start of each condition whether the sequences would consist of meaningful sentences or random lists. The 12 test trials were performed following completion of 2 practice sequences.

Each sequence was presented through headphones, at one word per second. To signal sequence completion, an abstract visual pattern and a non-verbal audio signal were presented following the last word in the sequence. In the recall conditions, participants were instructed to immediately recall as much of the sequence as possible, in the order in which it was presented. Responses were digitally recorded for transcription.

In the serial order recognition task, the cues at the end of the sequence were immediately followed by presentation of a test sequence, again at a rate of one word per second. In half the trials, the test sequence was identical to the first, whereas in the other half the sequence was altered so that two constituent words changed positions. Any word type at any serial position (except the first) could change, and positional changes always maintained meaningful grammatical structure (i.e. noun switched with noun, adjective with adjective, and verb with function word). Mean ‘swap distance’ (i.e. the number of positions across which the two words were exchanged) was equivalent for the sentence (3.21, SE = .21) and list (3.15, SE = .39) sets. Same order and different order trials were randomly intermixed.

Participants were instructed to judge whether the words in the re-presented sequence were in the same order as in the original sequence or if the order had changed, having been informed of the nature of the possible order change beforehand. Following completion of the second sequence, a same/different key press response was made. On responding ‘same’, the program moved on to the next trial. On responding ‘different’, the eight test words were vertically displayed on screen in the order of re-presentation, with a check box next to each word. Participants used the mouse to select the check boxes next to the two words they deemed to have switched positions, with the program automatically moving on when two were selected. No feedback was provided at any point.

Results

Data were analysed using standard and Bayesian repeated measures ANOVA and paired samples t-tests where appropriate, using JASP (JASP Team, 2017). For each analysis, we report the outcomes from the standard analyses, and the Bayes Factors (BF) indicating the strength of evidence for the inclusion of each effect within the relevant model, relative to the model without that effect. Finally, an analysis of chunking patterns in serial recall was also carried out (for this experiment, and Experiment 3), though for the sake of concision, these analyses are reported in an appendix.

Immediate Serial Recall

Each response was scored as correct if it was produced in the appropriate serial position[1]. Correct recall (as a proportion of the total presented sequence) of sentences and lists at each serial position is displayed in Figure 1a.

<Figure 1 about here>

A 2x8 repeated measures ANOVA revealed significant effects of sequence type, F (1,19) = 101.39, MSE = .04, p < .001, η^2 = .84, serial position (Greenhouse-Geisser corrected), F (3.75,71.27) = 87.24, MSE = .02, p < .001, η^2 = .82, and the interaction, F (7,133) = 6.52, MSE = .02, p < .001, η^2 = .26. Further analysis (Bonferroni-Holm corrected) revealed the sequence type effect to be significant at every serial position, though it was smaller at the first and second positions. A Bayesian ANOVA indicated the best model to contain sequence type, serial position, and the interaction (*BF* > 1000 vs. the null model), with Bayes Factors of >1000 supporting both main effects and the interaction.

Transposition Gradients and Error Analysis

Transposition gradients (e.g. Hurlstone et al., 2014; Nairne, 1992) were calculated to illustrate where in relation to the target position each item was recalled. These gradients are set out in Figure 1b, for sentences and lists, as a proportion of all responses, with target position centred at 0, anticipation errors at positions -7 to -1, and perseveration errors at positions 1-7. Examination of these gradients indicates that the sequence type manipulation impacted not only on correct responses, but also on the probability of producing words in incorrect positions in the response sequence. Increased positional uncertainty was apparent for list recall, with a relatively greater tendency for participants to erroneously recall words particularly just before their appropriate sequence position.

Examining order errors (i.e. recall of an item from the sequence in an incorrect sequence position) produced mean proportional rates of .19 (SE = .02) for sentences and .32 (.02) for lists, indicating a significantly higher order error rate for lists, *t* (19) = 6.12, *p* < .001, BF > 1000 (*Cohen's d* = 1.37). For omissions (i.e. failure to produce any response), we observed mean proportional rates of .05 (.01) for sentences and .13 (.02) for lists, a significant difference, *t* (19) = 5.43, *p* < .001, BF = 798 (*d* = 1.21). Finally, rates of intra-experiment intrusion errors (i.e. recall of an item from the experimental set that was not present on that trial) were .06 (.01) for sentences and .09 (.01) for lists, again a meaningful and significant difference, *t* (19) = 3.68, *p* = .002, BF = 25 (*d* = .82). Thus, sentential context led to improved recall order, and reductions in omissions and intrusions from the wider experimental set.

Serial recognition

Performance in the serial recognition task can be examined in terms of accuracy on the initial change detection task (combining change and no-change trials), and for those trials in which an order change was correctly detected, accuracy in the subsequent identification of which two words switched positions. Firstly, there was no significant difference in change detection accuracy (see Figure 2) for sentences and lists, t (19) = 1.07, p = .30, BF = .38, or 2.6 in favour of the null, (d = .24). However, accuracy in the identification of the words that had changed positions (following correct change detection) was .78 (.05) for sentences, and .60 (.05) for lists, a significant difference, t (19) = 3.25, p = .004, BF = 11 (d = .73). Therefore, while sentential structure only had a small and non-significant effect on the ability to detect changes in order, it did have a larger and significant effect on the accuracy of identifying precisely which words had changed positions in a sequence between presentations.

<Figure 3 about here>

Finally, and acknowledging the issues associated with such an analysis (e.g. differing chance rates), we directly compared recall and change detection recognition within a 2x2 repeated measures ANOVA (see Figure 2). There were significant effects of sequence type, F (1,19) = 31.70, MSE = .01, p < .001, η^2 = .63, and task, F (1,19) = 24.26, MSE = .02, p < .001, η^2 = .56, with mean recognition accuracy superior to recall. There was also an interaction between sequence type and task, F (1,19) = 14.80, MSE = .01, p = .001, η^2 = .44, with a substantially larger sentence superiority effect in recall (d = 2.25), relative to recognition (d = .24). A Bayesian ANOVA indicated the best model to contain sequence type, task, and the interaction (BF > 1000 vs. the null model), with Bayes Factors of >1000 supporting both main effects, and of 41 to 1 in support of the interaction.

Discussion

The recall advantage for constrained sentences over word lists found in previous studies (e.g. Baddeley et al., 2009) was again observed in this experiment. The consistency of this effect across recall accuracy, order errors, omissions, and intra-experiment intrusions suggests that the sentence advantage influences both item and order memory. We also report, for the first time, transposition gradients for sentences and lists. These illustrate how order errors tend to cluster around the correct serial position (Henson, Norris, Page, & Baddeley, 1996; Hurlstone et al., 2014), and appear relatively more likely to reflect anticipation errors, that is, the erroneously early recall of items, than postponements. The similarities between the shapes of the serial position curves and the gradients of transposition errors are surprising, and suggest that serial ordering mechanisms exert a similar effect on recall for both sentences and lists, consistent with a common underlying process or storage capacity (e.g. the episodic buffer).

In contrast to the large sentence effect in recall, we failed to find a significant sentence effect in the change detection phase of serial order recognition, with a Bayes Factor favouring the null model (i.e. without the inclusion of a sentence effect) by a factor of 2.6 to 1. This relative absence of an effect on aurally presented serial recognition is in line with outcomes of studies examining the effects of language knowledge at the individual word level (e.g. Gathercole et al., 2001; Macken et al., 2014). It might indicate that the sentence advantage as observed in serial recall emerges primarily during the retrieval process at output. These retrieval processes might also have been responsible for the sentence advantage that did emerge in the second stage of the recognition task, in which precise identification of the changed words was required. Alternatively, the absence of an effect might reflect the ability to make recognition judgments concerning auditory sequences using perceptual-based sequence matching (Macken et al., 2014), which could attenuate any effects that arise from stored language knowledge.

A possible alternative explanation of the apparent discrepancy between the substantial sentence effects observed in serial recall, and evidence supporting the absence of an effect in serial recognition, is that this reflects the relative insensitivity of the serial recognition methodology. The relatively meagre amount of data (1 data point per trial) that this binary correct/incorrect measure provides contrasts with the more data-rich method of serial recall. This discrepancy in potential sensitivity between response tasks may well have reduced our capacity to detect a genuine difference in recognition memory between sentences and lists. This may be particularly problematic when implemented within a limited number of trials all of which use the same sequence length for all participants, thus not allowing for intra-individual variation in performance levels.

Experiment 2

This experiment further explored whether the sentence effect emerges in serial recognition of aurally presented sequences. The methodology implemented in Experiment 1 may be potentially insensitive to the sequence-type (or indeed, any) manipulation, due to the relative paucity of data that two-alternative responses provide. This problem is often addressed in sensory psychophysics using the "staircase" or transformed up-down procedure (Levitt, 1971), a method that involves increasing the difficulty of the discrimination by one step whenever two successive correct responses are made and reducing it by the same amount following an error. In the case of serial recognition memory, each step corresponds to a change in list length. The reliability of the threshold can be increased by requiring more of these directional switches before terminating the trial, with each switch effectively serving as a within-trial replication. Experiment 2 therefore compared sentence and list recognition using the 'staircase' or transformed up-down span procedure. This method enables a focus on the sequence length at which each participant is accurate at an approximate rate of 70% (see Levitt, 1971), and provides a more sensitive measure of performance. If our previous failure to detect a sentence-list difference with auditory presentation is indeed due to lack of power or sensitivity using a simple change detection task within a limited number of set-length trials, then a replication using the staircase method should produce a reliable sentence advantage. In contrast, if the absence of a sentence advantage in Experiment 1 is genuine, evidence for the null effect should be observed again.

Method

Participants

Twenty undergraduate and postgraduate students (13 female) from the University of York took part, in return for course credit or financial payment.

Materials

Sequences were constructed using the item set, positional and order constraints, and audio files outlined in Experiment 1. Two sets of sentences and two sets of lists were created for each of the sequence lengths from 6-11 words, with 18 sequences in each set. Details of the content of sentences and lists at each sequence length are provided in the appendix. Of the 18 sequences at each length, 9 were selected for order change trials and 9 for no change trials.

Design and Procedure

All participants performed two blocks of each sequence condition. Blocks alternated between the two sequence types (e.g. sentences-lists-sentences-lists), with order counterbalanced across the experiment. Participants were informed at the start of each condition whether the sequences would consist of meaningful sentences or random word lists.

Each block started with 2 practice trials using sequences of 6 words, followed by the test set, which also started at length 6. Subsequent sequence length was then determined by change detection accuracy, using a transformed up-down procedure. Two successive correct responses led to sequence length increasing by 1 word on the following trial. Conversely, a single incorrect change detection response led to a reduction in sequence length by 1 word. A series of steps in one direction (constituting either a series of incorrect responses, or a series of correct responses) is termed a 'run'. A 'switch' is defined as a change in step direction, caused by two consecutive correct responses after an incorrect response (thus terminating a downward run), or an incorrect response after two correct responses (thus ending an upward run). The trial block continued until 8 'switches' in staircase direction were recorded. As such, the total number of trials varied between participants. If an incorrect detection response was made at length 6, condition difficulty remained at this level until two correct responses were made. At the upper length limit, sequence length remained at 11 words until an incorrect response was made.

The presentation and testing procedure within each trial was identical to the recognition method used in Experiment 1. Thus, each trial consisted of change detection followed (in the case of a "different" judgement) by change identification[3]. Sequences were randomly selected from the pool of 18 at each length, with each sequence only used once for each participant. There was an equal probability of change and no change trials being selected. Mean swap distance was equivalent for the sentence (2.84, SD = 1.46) and list (2.94, SD = 1.30) sets.

Results

Three measures of change detection accuracy were examined; proportion correct, mean sequence length at each switch point, and mean mid-run point (see Figure 3).

<Figure 3 about here>

For change detection proportion correct, change detection performance at length 11 was at chance (< .5) in both sequence conditions, and so was excluded from analysis. For lengths 6-10, mean proportion correct was .76 (SE = .04) for sentences and .66 (.04) for lists, a significant difference, t (19) = 3.32, p = .004, BF = 12 (d = .74).

The mean sequence length at each switch point represents the mean length of sequence that elicited each 'switch' (i.e. a run of two correct responses, or one incorrect response), with a longer mean length illustrating that participants had made more correct responses and advanced on to longer sequences within the staircase paradigm. Mean length per switch point was longer for sentences (8.48, SE = .23) than lists (7.82, SE = .17), a significant difference, t (19) = 4.02, p < .001, BF = 43 (d = .90).

Mean mid-run point score was calculated by taking the 8 switch points in a block, and for each one, obtaining the mean sequence length at which that switch occurred and the mean length of the switch preceding it (Levitt, 1971). For example, if a participant was incorrect at sequence lengths 10 and 9, and then produced two correct responses at length 8 (i.e. a 'switch'), this response 'run' would cover length 10 to length 8, with a 'mid-run point' score for this run of 9. By then averaging across the 8 switches, we can obtain a measure of the average midpoint in each run of responses (thus determining the sequence length at which participants were performing at approximately 70% correct). The mean mid-run point for sentences was 8.34 (SE = .21) while for lists it was 7.72 (.16), a significant difference, t (19) = 3.66, p = .002, BF = 23 (d = .82). The average mid-point of a run of responses (and therefore the sequence length at which detection was at 70%) occurred at a longer sequence length for sentences than lists, further indicating a sentence advantage in change detection.

Discussion

A sentence advantage was observed in the primary change detection measure of serial recognition in this experiment. This effect also emerged on the additional dependent variables that the staircase procedure provides (mean length per switch point and mean mid-point run). These findings contrast with evidence supporting a null sequence type effect from Experiment 1, and underlines the importance of optimizing task sensitivity when exploring whether different manipulations have a genuine effect on memory performance, thus minimising the probability of type 2 error. On this broader note, we would suggest that the staircase procedure appears to have substantial utility in a working memory context.

Based on this methodology, Experiment 2 suggests that the factors underlying the sentence advantage in memory are in fact not limited to response output, and are not due solely to processes such as redintegration operating during recall. Instead, the structure provided within meaningful sequences appears to be utilized earlier in the memory task, with impacts then emerging on subsequent memory judgments in tasks that do not have an explicit recall element.

Experiment 3

It was of interest to examine whether the patterns of findings using auditory presentation replicate when sequence presentation is visual. In our earlier serial recall experiments (Baddeley et al., 2009), comparable sentence effects were observed following both auditory and visual presentation of sequences. We would therefore predict such an effect to again emerge in the present study, and potentially show similar patterns across measures of accuracy, errors, and position uncertainty gradients as were observed in Experiment 1.

However, no previous studies have examined whether and how the sentence effect might emerge on serial recognition when sequences are encountered in the visual modality. Indeed, few studies have explored serial recognition using visual presentation. Macken et al. (2014) observed a lexicality effect on recognition when presentation was visual, in contrast to the absence of such effects using auditory presentation. Thus, it is apparent that a different mechanism may operate on serial recognition when presentation is auditory versus visual in nature, and that we should establish whether the same kind of effects emerge across modalities. Experiment 3 is therefore a replication of Experiment 1, extending to visual presentation. Immediate serial recall of constrained sentences and lists was assessed, with explorations of accuracy, error types, transposition gradients. Immediate serial recognition was implemented using a set number of trials and a single sequence length, as in Experiment 1[3].

Method

Participants

Twenty undergraduate and postgraduate students (13 female) from the University of York took part, receiving course credit or financial payment.

Materials

Sequences of 7 words were used in this experiment. Each sequence contained three nouns, two adjectives, one verb, and one function word. The sequence construction rules from the previous experiment were again implemented here.

Design and Procedure

The design and procedure was almost identical to that used in Experiment 1, with two exceptions. Firstly, all presentation was visual rather than auditory. Secondly, sequence length was reduced from 8 to 7 words. Each participant performed all 4 conditions (sentences vs. lists; recall vs. recognition), with 2 practice and 12 test sequences in each condition. Within the serial recognition task, mean swap distance was 2.57 (SE = .25) for sentences and 2.92 (.34) for lists.

Across all conditions, words were presented sequentially just above the centre of the screen in Arial 28-point font. As with auditory presentation in Experiment 1, presentation occurred at a rate of 1 word per second, each remaining on screen for 500ms, and separated by a 500ms delay.

Results

Immediate Serial Recall

Proportion correct[4] for each serial position is shown in Figure 4a. A 2x7 repeated measures ANOVA revealed significant effects of sequence type, *F* (1,19) = 78.43, *MSE* = .07, *p* < .001, η^2 = .81, serial position (Greenhouse-Geisser corrected), *F* (2.61,49.49) = 59.67, *MSE* = .03, *p* < .001, η^2 = 76, and the interaction, *F* (6,114) = 11.75, *MSE* = .01, *p* < .001, η^2 = .38. Further analysis (Bonferroni-Holm corrected) revealed the sequence type effect to be significant at every serial position, though it was smaller at the earlier positions. A Bayesian ANOVA indicated strongest support for the model containing both main effects and the interaction (*BF* > 1000 vs. the null).

<Figure 4 about here>

Transposition Gradients and Error Analysis

Transposition gradients were again calculated (centred with the target position at 0, and running from -6 to 6), and are illustrated in Figure 4b. As in Experiment 1, examination of these gradients clearly indicates that the sequence type manipulation impacted on correct responses, and on the probability of producing words in incorrect sequence positions.

Examination of order errors produced mean proportional rates of .10 (SE = .02) for sentences and .25 (.02) for lists, indicating a significantly higher order error rate for lists, *t* (19) = 6.37, *p* < .001, BF > 1000 (*d* = 1.42). For omissions, we observed mean proportional rates of .02 (.01) for sentences and .09 (.02) for lists, a significant difference, *t* (19) = 4.95, *p* < .001, BF = 309 (*d* = 1.11). Finally, rates of intra-experiment intrusion errors (i.e. recall of an item from the experimental set that was not present on that trial) were .09 (.01) for sentences and .13 (.02) for lists, a significant difference, *t* (19) = 2.20, *p* = .040, BF = 1.7 (*d* = .49). Thus, sentential context led to improved recall order, and reductions in omissions and intrusions from the wider experimental set (though the Bayes Factor for this latter outcome only indicates weak evidence for the effect).

Serial recognition

Change detection accuracy averaged at .86 (.02) for sentences, and .76 (.04) for word lists (see Figure 5). This difference was significant, t (19) = 2.47, p = .023, BF = 2.6, (d = .55). Accuracy in the identification of the words that had changed positions (following correct change detection) was .93 (.02) for sentences, and .77 (.05) for lists, a significant difference, t (19) = 3.92, p < .001, BF = 39, (d = .88). Thus, sentential structure had significant effects on the ability to detect changes in order, and the accuracy to identify the nature of these changes.

<Figure 5 about here>

Finally, we again directly compared recall and change detection recognition within a 2x2 repeated measures ANOVA (see Figure 5). There were significant effects of sequence type, F (1,19) = 50.80, MSE = .01, p < .001, η^2 = .73, and task, F (1,19) = 31.68, MSE = .02, p < .001, η^2 = .63, with mean recognition accuracy superior to recall. There was also an interaction between sequence type and task, F (1,19) = 13.38, MSE = .01, p = .002, η^2 = .41, indicating that the sentence advantage was larger for recall (d = 1.98) than for recognition (d = .55). A Bayesian ANOVA indicated the best model to contain sequence type, task, and the interaction (BF > 1000 vs. the null model), with Bayes Factors of >1000 supporting both main effects, and of 22 to 1 in support of the interaction.

Discussion

A large sentence advantage was again observed in immediate serial recall, this time using sequential visual presentation of each constituent word. As in previous studies, this effect appears to be both item- and order-based, with more order errors, omissions, and intra-experiment intrusions during recall of lists than sentences. Patterns of transposition gradients resembled those observed with auditory presentation in Experiment 1, with a higher rate of accurate responses for sentences at each point in the sequence, and position errors for both sequence types broadly indicating a locality constraint (Henson et al., 1996). As in Experiment 1, the similarities between serial position curves and transposition gradients point to the involvement of a common serial ordering mechanism in sentence and list recall.

A sentence advantage was also observed in both initial change detection and subsequent change identification. This contrasts with Experiment 1 using auditory presentation, where some evidence for a null effect of sequence type was observed, and is more in line with the outcomes using the Staircase method in Experiment 2. However, Bayes Factor analysis in the present experiment still only indicated 'anecdotal' evidence for a sentence advantage in serial recognition change detection (BF = 2.6). Therefore, a final experiment was carried out, implementing the Staircase method with serial recognition of visually presented sequences.

Experiment 4

The aims and methodology of the final experiment closely correspond to those of Experiment 2, but with sequence presentation changed to the visual modality. We wanted to examine whether the relatively weak evidence for a sentence advantage that was observed on the change detection measure of serial recognition in Experiment 3 using visual presentation is extended and strengthened using a more sensitive transformed up-down span methodology.

Method

Participants

There were 22 (7 male and 15 female) undergraduate and postgraduate students in this experiment, receiving course credit or financial payment for their participation.

Materials, Design, and Procedure

This experiment was essentially a replication of the staircase span method from Experiment 2, using the same materials from that study, but with visual presentation during encoding and test (see Experiment 3).

Results

Three measures of change detection accuracy were again examined; proportion correct, mean sequence length at each switch point, and mean mid-run point (see Figure 6). For proportion correct, as in Experiment 2, change detection performance on sequences of 11 words was at chance in both conditions, and so was excluded from analysis. Change detection proportion correct across lengths

6-10 was .76 (*SE* = .04) for sentences and .67 (.02) for lists (*d* = 1.05), a significant difference *t* (21) = 4.91, *p* < .001, BF = 357. For mean length per switch point, this was 8.19 (*SE* = .17) for sentences and 7.52 (.17) for lists (*d* = .89), a significant difference *t* (21) = 4.15, *p* < .001, BF = 72, (*d* = .81). Finally, the mean mid-run point for sentences was 8.19 (.17) while for lists it was 7.52 (.17), again a significant difference, *t* (21) = 4.13, *p* < .001, BF = 69 (*d* = .88).

<Figure 6 about here>

Cross-experiment comparison

Aside from presentation modality, identical methodologies were implemented in Experiments 2 and 4. It was therefore appropriate to combine the data from these experiments in a series of mixed 2x2 ANOVA, to identify whether the magnitude of the sentence advantage varied with modality of presentation at encoding and test. This produced large and significant effects of sentence type, but no effect of modality or interaction with sequence type, for all three outcome measures: Change detection (sequence type *F* (1,40) = 31.73, *MSE* = .01, *p* < .001, η^2 = .44; modality *F* (1,40) = .003, *MSE* = .01, *p* = .96, η^2 = .00; interaction *F* (1,40) = .03, *MSE* = .01, *p* = .86, η^2 = .00); Mean length per switch point (sequence type *F* (1,40) = 33.29, *MSE* = .03, *p* < .001, η^2 = .45; modality *F* (1,40) = .12, *MSE* = .02, *p* = .73, η^2 = .00; interaction *F* (1,40) = .003, *MSE* = .03, *p* = .96, η^2 = .00); Mean mid-run point (sequence type *F* (1,40) = 30.22, *MSE* = .03, *p* < .001, η^2 = .43; modality *F* (1,40) = .62, *MSE* = 1.04, *p* = .43, η^2 = .02; interaction *F* (1,40) = .03, *MSE* = .03, *p* = .86, η^2 = .00). Bayesian ANOVA in each case indicated the preferred model to contain sequence type only (*BF* > 1000 vs. null), with *BF* < 1 for modality and the interaction for all outcome measures.

Discussion

Experiment 4 outcomes resemble those of Experiment 2, showing a clear sentence advantage in the recognition task for proportion correct, length per switch point and mid-run point, this time with visual rather than auditory presentation. Thus, in contrast to the inconsistent and 'anecdotal' outcomes observed with the simple change detection procedure used in Experiments 1 and 3, the more sensitive staircase methodology used in Experiments 2 and 4 gives consistent results across modalities. Direct comparison of these effects across the two staircase experiments further illustrated the consistency of these effects, while indicating no variation as a function of presentation modality.

General Discussion

Across four experiments we provide new insights concerning how structured and unstructured verbal sequences might be handled in working memory. Examination of immediate serial recall in Experiments 1 (auditory presentation) and 3 (visual presentation) demonstrated a clear sentence advantage in terms of correct performance and positional certainty across sequence positions, item and order error rates, as well as chunk size and order. In contrast, these experiments produced mixed and equivocal impacts of sequence structure on serial recognition performance. However, when recognition was examined using a staircase or transformed up-down procedure in Experiment 2 and 4, consistent and reliable sentence effects then emerged, across auditory and visual modalities.

Effects of sentence structure previously observed on immediate serial recall (Baddeley et al., 2009) were replicated and extended in the present study. This effect appears to impact on both item and order information. Examination of transposition gradients shed further light on how items are recalled across the different sequence positions. For both the auditory and visual modalities, sentential structure appears to reduce positional uncertainty and facilitate the recall of each presented item in its appropriate position. Both sentences and lists show something of a locality constraint (Henson et al. 1996) with reduced probability of displacement of list items to more distant output positions. List recall appears to be less precise in its positional certainty, and it is primarily at these adjacent positions that order errors (either representing anticipations or perseverations) can be observed. Thus, despite large differences in absolute levels of recall, serial position curves and distributions of transposition errors for sentences and word lists have very similar properties. This parallelism is clear and is more in line with common processes underpinning performance in both cases, as in the episodic buffer account (Baddeley et al., 2009, 2011), than with the idea that memory for sentences and lists involves fundamentally different representations, as in the conceptual regeneration account (Potter & Lombardi, 1998). However, such an interpretation must be viewed with caution as serial order may be a common problem solved in the same way within different systems, and it is also possible that our use of constrained sentences exaggerates similarities with word lists.

We also considered whether sentential structure might provide a useful source of support for processes of retrieval-based redintegration (Saint-Aubin & Poirier, 2000) that could operate at both the single-word and multi-word sequential level. Thus, when words are embedded into sentences at presentation, the structure and redundancy this provides might prove useful in cueing appropriate responses and inhibiting incorrect responses that might be grammatically or conceptually incorrect. We therefore extended the current exploration to serial recognition, a task that places relatively reduced demands on active output, and in which selection and redintegration of items in memory for retrieval is unlikely to play a major role. The equivocal outcomes observed using a more traditional methodological approach, with anecdotal evidence for the null model in Experiment 1 (auditory presentation) and for the presence of an effect in Experiment 3 (visual presentation) might lead us to conclude that the sentence advantage in serial recall is primarily output-based. Alternatively, these results at least somewhat parallel the findings of Macken et al. (2014), who observed lexicality effects in the visual but not the auditory modality, and attributed this to the application of global-level perceptual matching in the auditory modality and item segmentation arising from verbal recoding in the visual. Our observations of somewhat contrasting sequence-level linguistic effects (i.e. the sentence advantage) in Experiments 1 and 3 would on the face of it fit with these claims, although the strength of evidence (as indicated by Bayes Factor analysis) is weak in each case.

However, these possible explanations are undermined by stronger and more consistent evidence for sentence effects on serial recognition using a more sensitive up-down staircase measure (Experiments 2 and 4; see Figure 7 for a comparison of effect sizes). These outcomes highlight the importance of optimising task sensitivity to avoid type 2 errors. The simple change detection method (using a set length and a limited number of trials) used in Experiments 1 and 3 is widely featured in memory studies, but has the drawback that each present-absent decision trial conveys limited information, thus requiring a larger number of data points for reliable results. The insensitivity of this task has also been commented upon by Jefferies et al. (2006) whose evidence suggested this might account for the apparent discrepancy between their own findings of a small lexicality effect and those of Gathercole et al. (2001) reporting none. Thus, the staircase method, used extensively in sensory psychophysics, should perhaps be employed more frequently in working memory research. If each switch provides a within-trial replication, reliability can therefore be increased by specifying a greater number of reversals before concluding a trial. The task is also useful in identifying the optimal level of sensitivity at an individual participant level, thus providing more data points at sensitive performance levels and reducing trial redundancy. It is thus likely to prove a more robust and reliable measure of the underlying memory function, and might be adopted more widely in working memory research.

<Figure 7 about here>

Given the emergence of strong evidence for a sentence superiority effect using this methodology, we suggest that sentence structure benefits both recognition and recall through strengthening of representations during encoding and storage. How might these processes be captured within working memory? When verbal sequences are initially encountered, this input interacts with different levels (e.g. conceptual, lexical, syntactic and grammatical) of stored language knowledge held in long-term memory, thus activating representations of meaning and structure. Different theoretical perspectives would then provide distinct accounts for how the sentence superiority effect manifests. One possibility is that syntactic and conceptual binding processes stabilise the representation of information about individual words and word order in memory, analogous to the way lexical binding processes have been assumed to stabilise phonological representations of items (see also Majerus, 2013 for a similar proposal). The outputs of this processing might then be represented as bound elements in a modality-general storage capacity such as the episodic buffer (Baddeley, 2000). As with simple forms of visual binding (Allen, Baddeley, & Hitch, 2006, 2014), or the utilisation of meaningful verbal-spatial associations (Calia, Darling, Havelka, & Allen, under review), this would appear to operate in a relatively automatic fashion (Baddeley et al., 2009).

While this approach is intended to capture a range of phenomena, other theoretical accounts have been proposed to more specifically handle mnemonic processing of meaning and linguistic structure. For example, Potter and Lombardi (1990, 1998; Lombardi & Potter, 1992; see also Rummer & Engelkamp, 2001, 2003; Schweppe et al., 2011) propose that sentence meaning is extracted at a conceptual level during comprehension, which then drives regeneration of the sequence via persisting lexical, syntactic, and phonological priming. This extraction of conceptual information from the point of encoding might predict a sentence superiority effect across different tasks, as unstructured list memory, in contrast, is dependent on preservation of a surface phonological representation. Alternatively, the sentence superiority effect may reflect more effective storage within activated long-term memory. For example, MacDonald (2016) has recently described how utterance planning might underlie maintenance and ordering in verbal memory tasks using either structured sentences or random word lists, characterising these utterance action plans as activated portions of long-term memory under the focus of attention (Cowan, 2005). Taking a different approach, Ericsson and Kintsch (1995) developed the concept of ‘long-term working memory’, suggesting that long-term memory can be used as a form of temporary storage, provided retrieval cues are actively retained within ‘short-term working memory’ itself. In the case of text comprehension, incoming text is processed via interactions between working and long-term memory, with the end products of this multi-level construction and segment integration process retained in LTM. Each of these theoretical approaches might plausibly predict sentence superiority effects to emerge across serial recall and recognition tasks. It would be useful for future work to examine which might provide the most parsimonious explanation for such effects.

It is also worth noting that the sentence superiority effect, while somewhat larger and more consistent using the staircase methodology in Experiments 2 and 4, was nevertheless substantially smaller than that observed in serial recall (see Figure 7). Regardless of presentation modality, serial recall yields sentence superiority effect sizes of $d \approx 2,0$, classed by Sawilowsky (2009) as ‘huge’. In contrast, recognition measures produce small-medium effects ($d \approx 0.2$-0.5) using set length, and large effects ($d \approx 0$ .8) within the up-down staircase procedure. Notwithstanding broader measurement differences between these tasks, these disparate effect sizes might indicate the sentence superiority effect to have an output-based component. Candidate mechanisms that might capture this include redintegrative processes operating on degraded representations at recall, regeneration of the sequence via lexical, syntactic, and phonological priming, based on conceptual representations formed during encoding (Potter & Lombardi, 1990, 1998), or the implementation of utterance plans for recall (MacDonald, 2016). In each case, such processing would be assumed to be qualitatively or quantitatively superior for structured sequences, relative to word lists, and would be emphasized in serial recall and minimized in serial recognition. It is possible, furthermore, that such retrieval-based processing may also explain outcomes observed in recognition. Indeed, it is likely that recall and recognition lie on a continuum regarding the retrieval demands involved in each task, with recognition reducing but not eliminating the impacts of retrieval-based processing. It could be argued that the change detection task might involve an initial stage of covertly recalling the remembered sequence for comparison with the test sequence. While it is not possible to strongly rule out such an explanation based on the current work, it is perhaps unlikely to be an effective strategy given that re-presentation of the sequence immediately followed completion of the original sequence. Furthermore, if covert recall were an obligatory stage in decision phase of the recognition task, we would expect recall accuracy to be at least as high as recognition, when in fact we observed the classic finding that recognition was superior (McDougall, 1904). Therefore, we would argue that retrieval-based effects of sentence structure emerge in addition to those occurring upstream, during initial encoding.

In conclusion, it is apparent that sentence superiority effects in immediate serial recall emerge across multiple forms of analysis, including item recall accuracy, ordering, and transposition gradients. We also show that it produces sizeable effects in serial recognition, a task that minimizes explicit recall/output demands, though observation of such effects on this form of task are partly dependent on using a sufficiently sensitive paradigm. The magnitude of such effects, however, remains substantially larger in recall, relative to recognition. These outcomes generally remain consistent across presentation modality. Overall, the results indicate that sentence structure beneficially impacts across encoding, storage, and retrieval, rather than being localized at any one phase of the task.

References

Allen, R. J., & Baddeley, A. D. (2008). Working memory and sentence recall. In A. Thorn, & M. Page (Eds.), *Interactions between short-term and long-term memory in the verbal domain* (pp. 63–85). Hove: Psychology Press.

Allen, R. J., Baddeley, A. D., & Hitch, G. J. (2006). Is the binding of visual features in working memory resource-demanding? *Journal of Experimental Psychology: General*, *135*(2), 298-313.

Allen, R. J., Baddeley, A. D., & Hitch, G. J. (2014). Evidence for two attentional components in visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(6), 1499-1509.

Baddeley, A. D. (1966a). The influence of acoustic and semantic similarity on long-term memory for word sequences. *The Quarterly Journal of Experimental Psychology*, *18*(4), 302-309.

Baddeley, A. D. (1966b). Short-term memory for word sequences as a function of acoustic, semantic and formal similarity. *The Quarterly Journal of Experimental Psychology*, *18*(4), 362-365.

Baddeley, A. D., Allen, R. J., & Hitch, G. J. (2011). Binding in visual working memory: The role of the episodic buffer. *Neuropsychologia*, *49* (6), 1393-1400.

Baddeley, A. D., Hitch, G. J., & Allen, R. J. (2009). Working memory and binding in sentence recall. *Journal of Memory and Language, 61*, 438-456.

Brener, R. (1940). An experimental investigation of memory span. *Journal of Experimental Psychology, 26*, 467-483.

Burgess, N., & Hitch, G. (2005). Computational models of working memory: Putting long-term memory into context. *Trends in Cognitive Sciences, 9,* 535-541. doi:10.1016/j.tics.2005.09.011

Calia, C, Darling, S., Havelka, J., & Allen, R.J. (under review). Visuospatial bootstrapping: binding useful visuospatial information during verbal working memory encoding does not require set-shifting executive resources.

Darling, S., Allen, R. J., & Havelka, J. (2017). Visuospatial Bootstrapping: When Visuospatial and Verbal Memory Work Together. *Current Directions in Psychological Science*, *26*(1), 3-9.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological review*, *102*(2), 211-245.

Gathercole, S. E., Pickering, S., Hall, M., & Peaker, S. (2001). Dissociable lexical and phonological influences on serial recognition and serial recall. *Quarterly Journal of Experimental Psychology, 54A*, 1-30.

Henson, R. N., Norris, D., Page, M., & Baddeley, A.D. (1996). Unchained memory: Error patterns rule out chaining models of immediate serial recall. *The Quarterly Journal of Experimental Psychology: Section A*, *49*(1), 80-115.

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

Hurlstone, M. J., Hitch, G. J., & Baddeley, A. D. (2014). Memory for serial order across domains: An overview of the literature and directions for future research. *Psychological bulletin*, *140*(2), 339-373.

Jacobs, J. (1887). Experiments on "prehension". *Mind*, *12,* 75-79.

Jefferies, E., Frankish, C., & Lambon Ralph, M. A. (2006). Lexical and semantic influences on item and order memory in immediate serial recognition: Evidence from a novel task. *Quarterly Journal of Experimental Psychology, 59*, 949-964.

Kendall, M. (1938). A New Measure of Rank Correlation. *Biometrika 30*, 81–89.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America, 33,* 467–476

Lombardi, L., & Potter, M. C. (1992). The regeneration of syntax in short term memory. *Journal of Memory and Language*, *31*(6), 713-733.

JASP Team (2017). JASP (version 0.8.1.2)[Computer software].

MacDougall, R. (1904). Recognition and recall. *The Journal of Philosophy, Psychology and Scientific Methods*, *1*(9), 229-233.

MacDonald, M. C. (2016). Speak, Act, Remember: The Language-Production Basis of Serial Order and Maintenance in Verbal Memory. *Current Directions in Psychological Science*, *25*(1), 47-53.

Macken, B., Taylor, J. C., & Jones, D. M. (2014). Language and short-term memory: The role of perceptual-motor affordance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(5), 1257-1270.

Majerus, S. (2013). Language repetition and short-term memory: an integrative framework. *Frontiers in Human Neuroscience.* doi: 10.3389/fnhum.2013.00357

Nairne, J. S. (1992). The loss of positional certainty in long-term memory. *Psychological Science*, *3*(3), 199-202.

Potter, M. C., & Lombardi, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory and Language*, *29*(6), 633-654.

Potter, M. C., & Lombardi, L. (1998). Syntactic priming in immediate recall of sentences. *Journal of Memory and Language*, *38*(3), 265-282.

Rummer, R., & Engelkamp, J. (2001). Phonological information contributes to short-term recall of auditorily presented sentences. *Journal of Memory and Language*, *45*(3), 451-467.

Rummer, R., & Engelkamp, J. (2003). Phonological information in immediate and delayed sentence recall. *The Quarterly Journal of Experimental Psychology: Section A*, *56*(1), 83-95.

Saint-Aubin, J., & Poirier, M. (2000). Immediate serial recall of words and nonwords: Tests of the retrieval-based hypothesis. *Psychonomic Bulletin & Review*, *7*(2), 332-340.

Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods, 8(2),* 597 – 599.

Schweickert, R. (1993). A multinomial processing tree model for degradation and redintegration in immediate recall. *Memory & Cognition, 21*, 168-175.

Schweppe, J., & Rummer, R. (2007). Shared representations in language processing and verbal short-term memory: The case of grammatical gender. *Journal of Memory and Language*, *56*(3), 336-356

Schweppe, J., Rummer, R., Bormann, T., & Martin, R. C. (2011). Semantic and phonological information in sentence recall: Converging psycholinguistic and neuropsychological evidence. *Cognitive Neuropsychology*, *28*(8), 521-545.

Tulving, E., & Patkau, J. E. (1962). Concurrent effects of contextual constraint and word frequency on immediate recall and learning of verbal material. *Canadian Journal of Psychology, 16*, 83-95.

Footnotes

[1] Note that we also scored performance using the adjacency method as reported by Baddeley et al. (2009). A 2x8 repeated measures ANOVA revealed significant effects of sequence type, F (1,19) = 139.49, MSE = .02, p < .001, η^2 = .89, serial position, F (7,133) = 51.58, MSE = .02, p < .001, η^2 = .73, and the interaction, F (7,133) = 9.79, MSE = .01, p < .001, η^2 = .34. The sentence effect was significant at every serial position, though it was smaller at the first and final positions, relative to positions 2-7. Proportion correct, collapsed across serial positions, was .77 (std error = .03) for sentences and .57 (.03) for lists (*Cohen's d* = 2.64).

[2] Experiments 2 and 4 were focused on change detection performance rather than the secondary measure of change identification. In addition, the staircase method meant that number of trials at each length was determined by change detection performance and was not balanced across sequence types. Therefore, change identification is not reported or analysed further, although outcomes on this measure replicated those observed in the first two experiments.

[3] Experiment 3 was carried out before Experiment 2, but for ease of exposition the order in which the experiments are reported has been changed.

[4] Using the adjacency scoring method, a 2x7 repeated measures ANOVA revealed significant effects of sequence type, F (1,19) = 43.21, MSE = .08, p < .001, η^2 = .70, serial position, F (6,114) = 39.98, MSE = .02, p < .001, η^2 = .68, and the interaction, F (6,114) = 8.95, MSE = .01, p < .001, η^2 = .32. The sequence type effect was significant at every serial position, though it was smaller at the earlier positions. Proportion correct, collapsed across serial positions, was .80 (std error = .03) for sentences and .58 (.04) for lists (d = 1.47).

Figure 1. a) Proportion correct (and standard error) in serial recall for sentences and lists as a function of serial position, and b) Transposition gradients in serial recall for sentences and lists in Experiment 1.

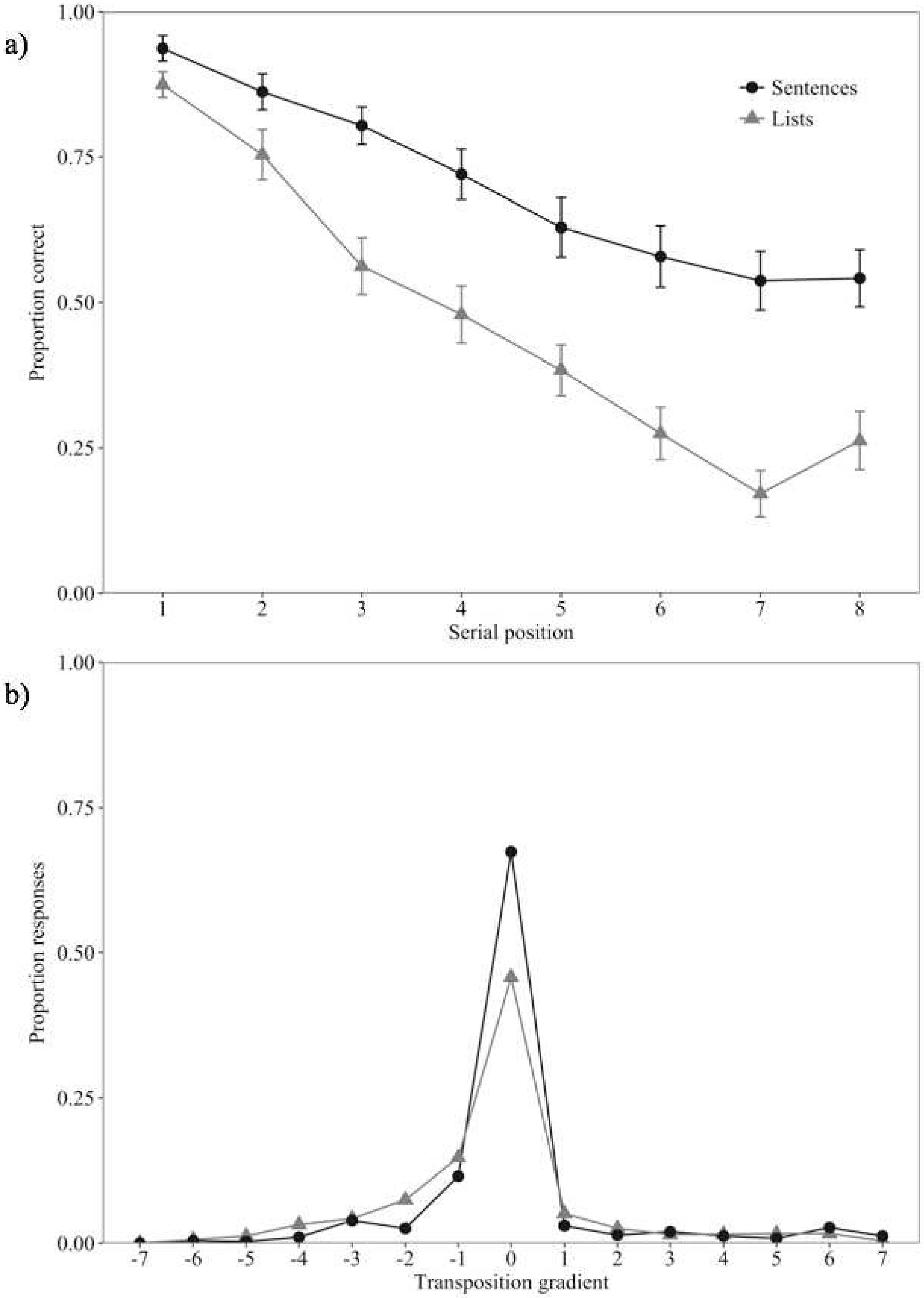


Figure 2. Mean performance in Experiment 1 for serial recall and recognition. Error bars show standard error.

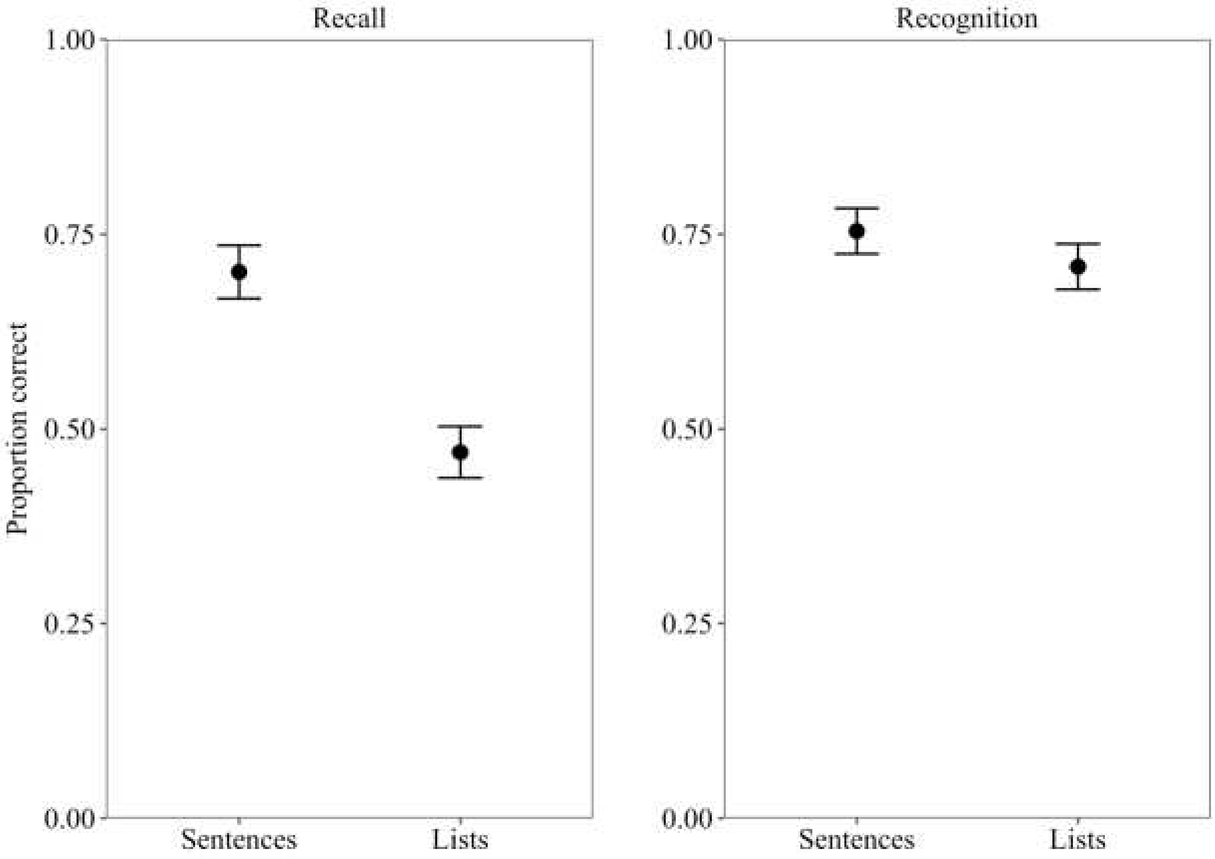


Figure 3. Recognition proportion correct, switch point length, and mid-run point in Experiment 2. Error bars show standard error.

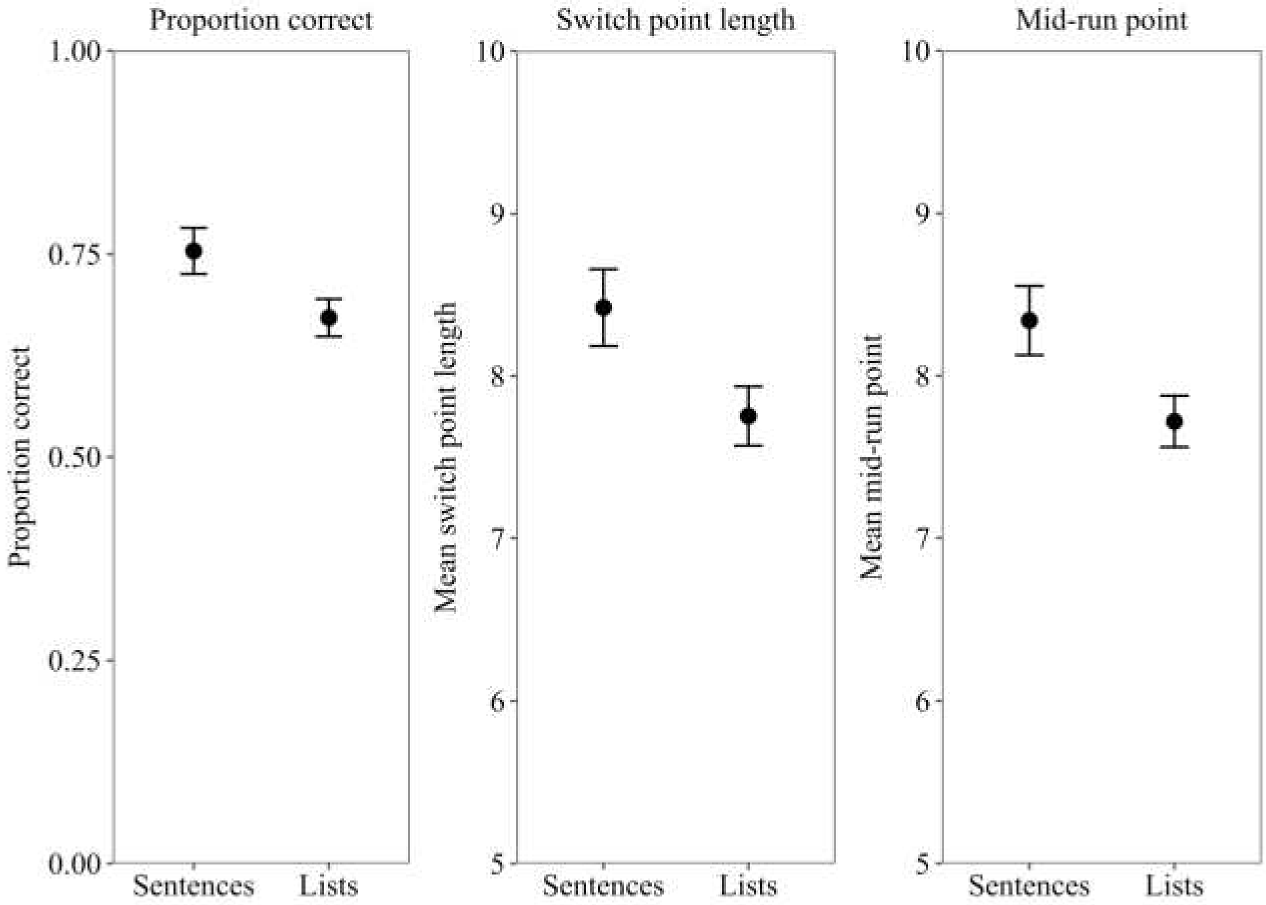


Figure 4. a) Proportion correct (and standard error) in serial recall for sentences and lists as a function of serial position, and b) Transposition gradients in serial recall for sentences and lists in Experiment 3.

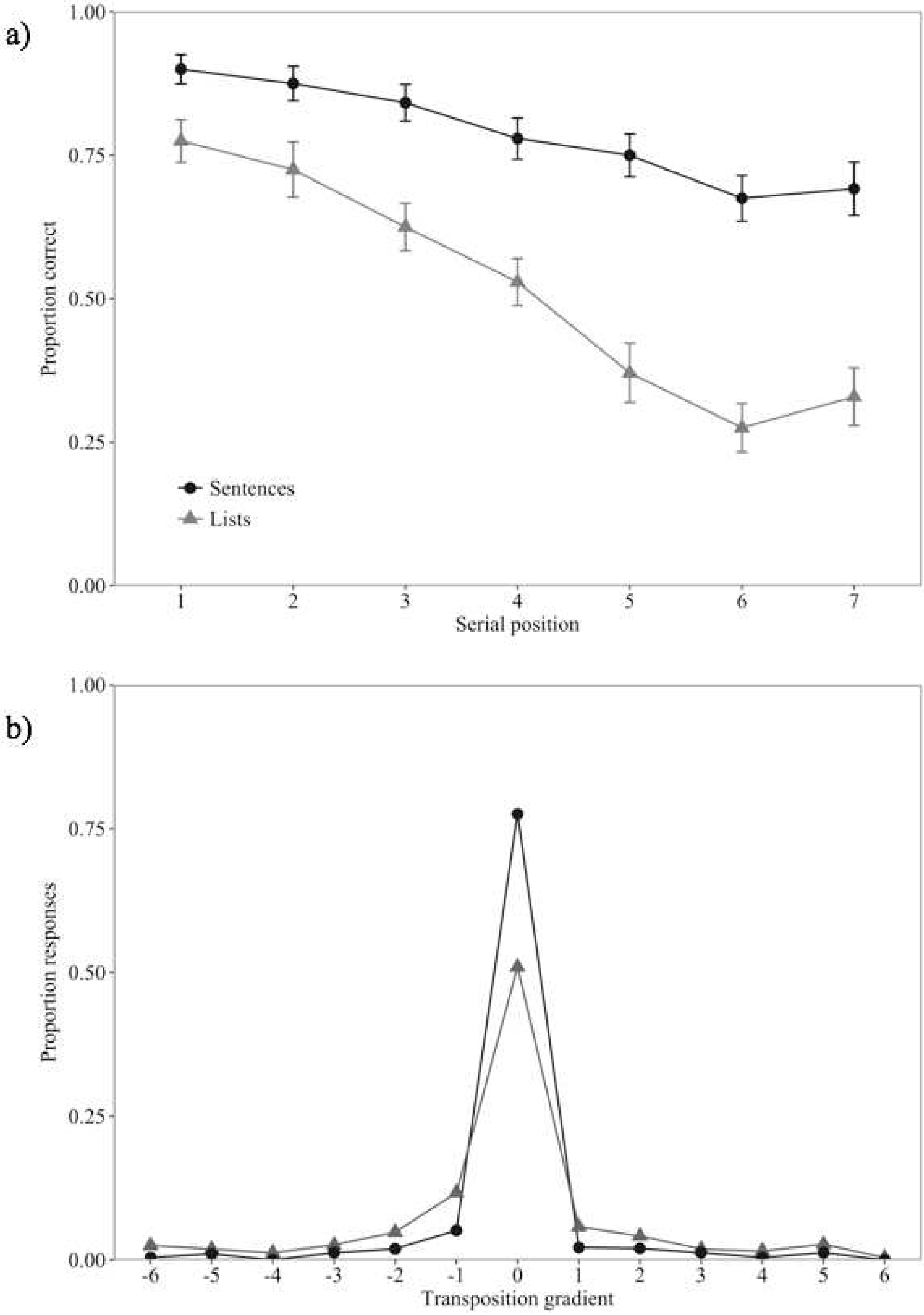


Figure 5. Mean performance in Experiment 3 for serial recall and recognition. Error bars show standard error.

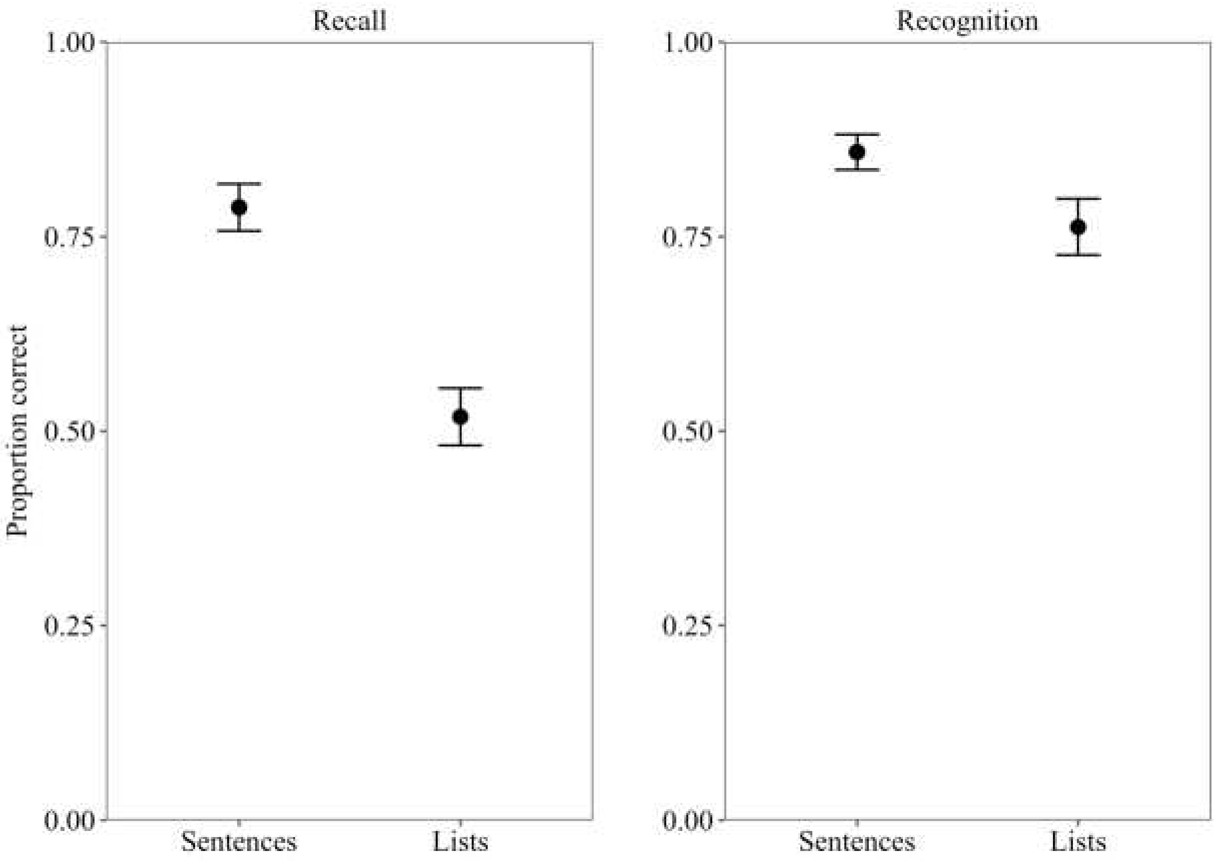


Figure 6. Recognition proportion correct, switch point length, and mid-run point in Experiment 4. Error bars show standard error.

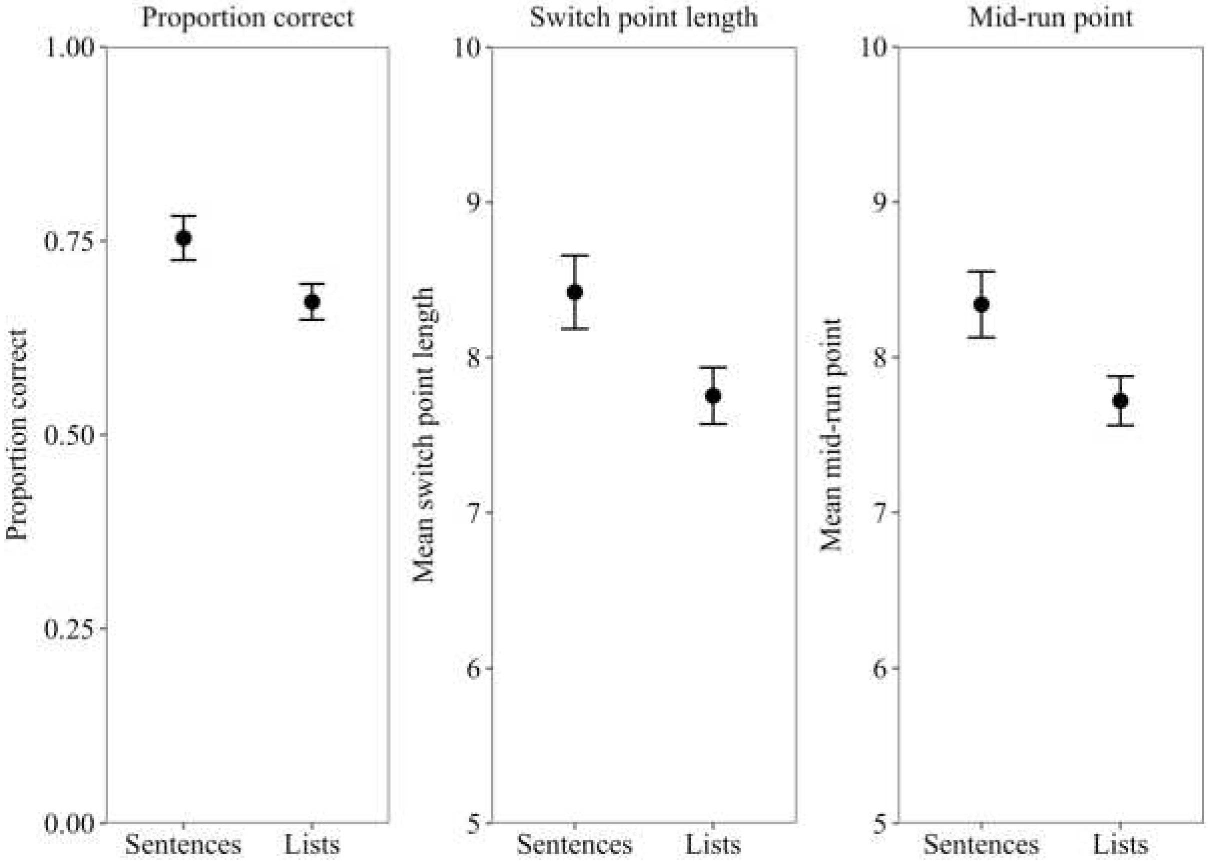
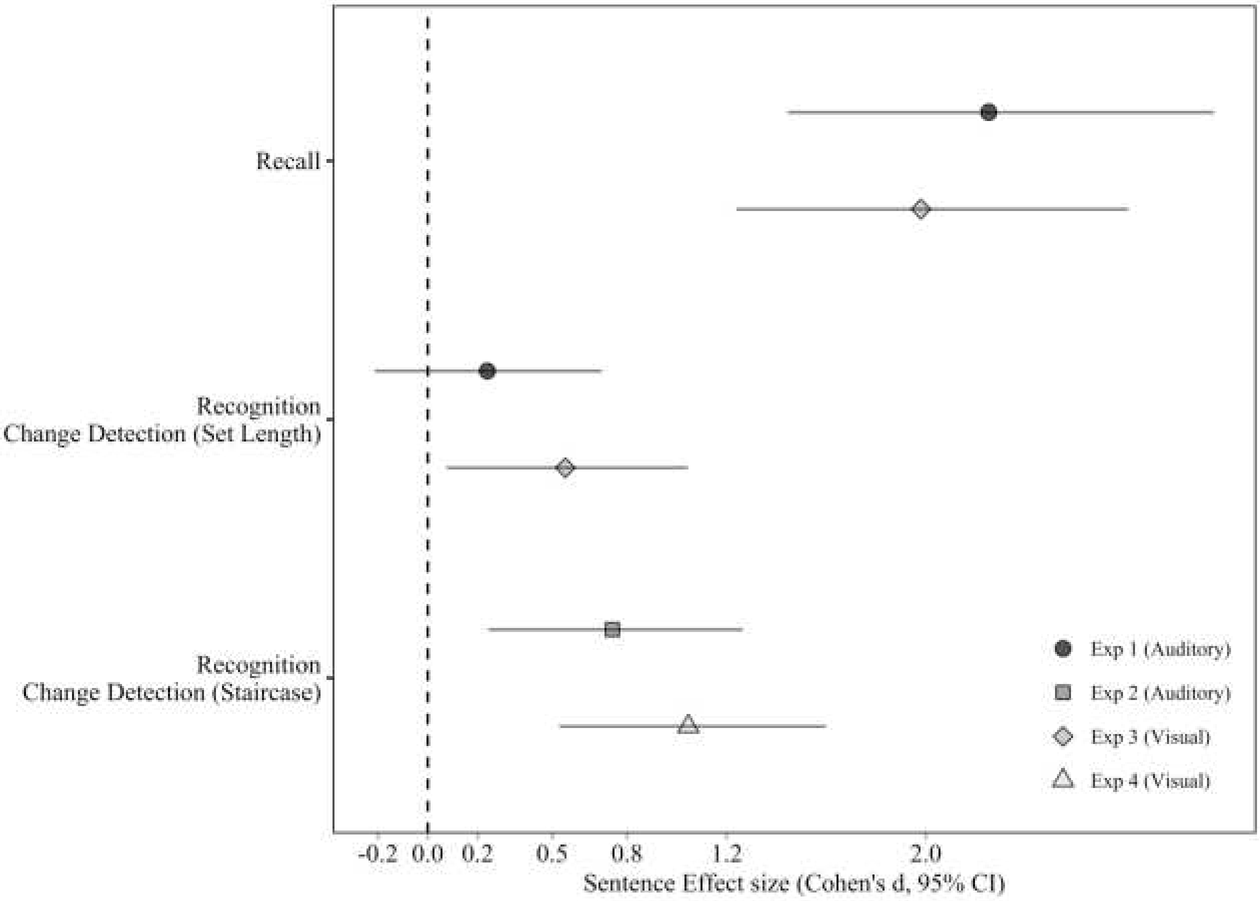


Figure 7. Sentence superiority effect sizes (Cohen's d) observed on each of the primary outcome measures in Experiments 1-4



Appendix: Chunking analyses

Tulving and Patkau (1962) found that participants produced larger ‘adopted chunks’ (groups of items recalled in the correct sequential order) when recalling text passages with a relatively higher approximation to English. Our previous exploration of sentence memory (Baddeley et al., 2009) did not allow examination of this issue or application of such methods, as we used unbalanced sequence lengths in most of those experiments. The present use of equal sequence lengths therefore enables us to examine ‘adopted chunking’ within the constrained sentence methodology. In line with Tulving and Patkau (1962), we classed a recalled ‘chunk’ as an unbroken sequence of words recalled in the order in which they were originally presented. This provides a measure of the mean size of recalled chunks. In addition, we also examined whether these recalled chunks were themselves produced in the correct relative order, using Kendall’s tau (τ) rank correlation coefficient (Kendall, 1938). This method can be used to calculate the number of pairs of chunks that are in the correct or incorrect relative order across the response sequence, as a function of the total number of chunks recalled. It is expressed as:

$$\tau = \frac{n_c - n_d}{\frac{1}{2} n\,(n\text{-}1)}$$

where n_c is the number of correctly ordered chunk pairs, n_d is the number of incorrectly ordered chunk pairs, and n is the total number of chunks. So, for example, if the sequence ABCDEF was recalled as “AB D C EF”, this would be scored as 4 chunks, with 5 pairs in the correct relative order (AB-D, AB-C, AB-EF, D-EF, C-EF) and 1 pair in the incorrect relative order (D-C), giving a τ score of .67.

In Experiment 1, the mean size of recalled chunks was 4.04 words (.34) for sentences and 2.43 words (.21) for lists, a significant difference, t (19) = 7.44, p < .001, BF > 1000 (d = 1.66). The Kendall’s tau analysis of chunk order produced a score of .75 (.03) for sentences and .69 (.02) for lists. This difference was significant, t (19) = 2.24, p = .037, BF = 1.8 (d = .50).

In Experiment 2, mean size of recalled chunks was 4.53 words (.30) for sentences, compared to 2.62 words (.25) for lists, a significant difference, *t* (19) = 6.50, *p* < .001, BF > 1000, (*d* = 1.45). The Kendall's tau analysis of correct order of chunk recall was .83 (.04) for sentences and ,66 (.04) for lists. Again, this difference between sequence types was significant, *t* (19) = 4.02, *p* < .001, BF = 48 (*d* = .90).

Thus, for both modalities, presentation of items within a sentence-like structure led to recall of larger chunks, in a more appropriate order, indicating that ordering mechanisms are important at the individual item level, within multi-word chunks, and across multiple items, between chunks. We acknowledge that this form of chunking analysis is directly based on recall performance, and is therefore not independent of recall accuracy. As such, it is not possible to confidently assert whether this represents chunking during encoding, or the retrieval of chunk-like segments at recall. In either case though, these analyses are instructive in indicating how structure impacts on performance.