Can the effects of temporal grouping explain the similarities and differences between free recall and serial recall?

Jessica Spurgeon[a], Geoff Ward[a], William J. Matthews[b] & Simon Farrell[c]

[a] University of Essex

[b] University of Cambridge

[c] University of Western Australia

Correspondence to:

Jessica Spurgeon or Geoff Ward,

Department of Psychology,

University of Essex,

Wivenhoe Park,

Colchester, Essex

CO4 3SQ, UK

Telephone: +44 1206 873799 or +44 1206 873800

Fax: +44 1206 873590

E-mail: jsmithy@essex.ac.uk or gdward@essex.ac.uk

Running header: Temporal grouping effects in immediate memory

Abstract

Temporal grouping can provide a principled explanation for changes in the serial position curves and output orders that occur with increasing list length in immediate free recall (IFR) and immediate serial recall (ISR). To test these claims, we examined the effects of temporal grouping on the order of recall in IFR and ISR of lists of between 1 and 12 words. Consistent with prior research, there were significant effects of temporal grouping in the ISR task with mid-length lists using serial recall scoring, and no overall grouping advantage in the IFR task with longer list lengths using free recall scoring. In all conditions, there was a general tendency to initiate recall with either the first list item or with one of the last four items, and then to recall in a forward serial order. In the grouped IFR conditions, when participants started with one of the last four words, there were particularly heightened tendencies to initiate recall with the first item of the most recent group. Moreover, there was an increased degree of forward-ordered transitions within-groups than across groups in IFR. These findings are broadly consistent with Farrell's (2012) model---in which lists of items in immediate memory are parsed into distinct groups and participants initiate recall with the first item of a chosen cluster---but also highlight shortcomings of that model. The data support the claim that grouping may offer an important element in the theoretical integration of IFR and ISR.

242 words

Keywords: working memory, free recall, serial recall, grouping, clustering

Immediate free recall (IFR) and immediate serial recall (ISR) are two widely-used and theoretically-important immediate memory tasks that have been highly influential in the development of accounts of short-term memory (e.g., Atkinson & Shiffrin, 1971; Glanzer, 1972) and working memory (e.g., Baddeley, 1986; Baddeley & Hitch, 1974), respectively. The over-arching aim of this paper is to explore whether temporal clustering can provide a principled explanation for the observed similarities and differences across the two tasks, thus offering important constraints on potential theoretical integration of these different research domains.

In tests of IFR (e.g., Murdock, 1962), participants are typically presented with a series of 10-40 words one at a time; at the end of the list, participants try to remember as many of the words as they can, and are free to recall these words in any order that they wish. In such tests, participants tend to show (1) enhanced recall of the most recent items, *the recency effect*, which is often attributed to the direct output of the contents of short-term memory, and (2) enhanced recall of the earliest list items, *the primacy effect*, which is often attributed to the strengthening of associations involving these words in long-term memory following their selective rehearsal (e.g., Rundus, 1971). Early accounts proposing a distinction between short-term (or primary) memory and long-term (or secondary) memory relied heavily on data from IFR (e.g., Atkinson & Shiffrin, 1971; Waugh & Norman, 1965).

In tests of ISR (e.g., Crannell & Parrish, 1957; Miller, 1956), participants are typically presented with shorter lists of 5-8 items; at the end of the list, participants are required to recall the items in the same serial order as they had been presented. Early studies recognised a capacity limit, referred to as the memory span, which refers to the maximum number of items that could be repeated back exactly in the same order on half the trials. Capacity limits have provided important empirical evidence for understanding short-term (Broadbent, 1975; Miller, 1956) and working memory (e.g., Cowan, 2000, 2005). The fact that the memory span

was later found to be sensitive to the phonological similarity (Baddeley, 1966) and the syllable length of the words in the list (Baddeley, Thomson & Buchanan, 1975) has also been central in underpinning the proposed Phonological Loop component of working memory (e.g., Baddeley, 1986; Baddeley & Hitch, 1974).

Given such an illustrious history, one might imagine that contemporary accounts of short-term memory and working memory might be well placed to explain performance in both ISR and IFR, especially given the similarity of the two tasks. However, somewhat surprisingly, there are many influential theoretical accounts proposed to explain only ISR performance (Baddeley, 1986, 2012; Botvinick & Plaut, 2006; Brown, Preece & Hulme, 2000; Burgess & Hitch, 1999, 2006; Farrell & Lewandowksy, 2002; Henson, 1998; Lewandowsky & Farrell, 2008; Page & Norris, 1998, 2003) and many other influential theoretical accounts proposed to explain only IFR performance (Atkinson & Shiffrin, 1971; Davelaar, Goshen-Gottstein, Ashkenazi, Haarmann, & Usher, 2005; Howard & Kahana, 1999, 2002; Raaijmakers & Shiffrin, 1981; Tan & Ward, 2000). While some influential accounts of working memory (e.g., Chen & Cowan, 2009; Cowan, 1988, 1999, 2000, 2005) focus on explaining capacity limitations across a wide range of tasks, these theories have less to say about the precise patterns of recall---such as serial position functions, output orders, and error data---that have been instrumental in the development and testing of theories of working memory.

As discussed by Ward, Tan, and Grenfell-Essam (2010), one difficulty for an integrated account of ISR and IFR is that, in line with the canonical patterns of data observed in the respective tasks, theories of ISR seek to explain large primacy effects and far reduced recency effects, whereas theories of IFR seek to explain large recency effects and far reduced primacy effects. However, recent empirical evidence suggests that the similarities of the two tasks substantially outweigh the differences when comparisons are made using the same list

length and scoring systems (e.g., Bhatarah, Ward, Smith & Hayes, 2009; Bhatarah, Ward & Tan, 2008; Grenfell-Essam & Ward, 2012; Spurgeon, Ward & Matthews, 2014; Ward *et al.*, 2010). At short list lengths, participants in both tasks often initiate recall with the first list item and then proceed in forward serial recall. That is, when participants are asked to recall *in any order* "mouse hat dog stairs" they tend to output "mouse hat dog stairs" even though there was no order requirement in the task instructions. Similarly, with longer lists, participants in ISR often find it difficult to recall the start of the list and instead initiate recall with a sequence of end-of-list items (Ward *et al.*, 2010). Collectively, these recent findings have encouraged interest in greater unification between theories of short-term memory and longer-term memory (e.g., Anderson, Bothell, Lebiere, & Matessa, 1998; Brown, Chater, & Neath, 2008; Brown, Neath, Chater, 2007; Farrell, 2012; Grossberg & Pearson, 2008; Hurlstone, Hitch & Baddeley, 2014; Kahana, 2012b).

In this paper, we examine whether temporal grouping might explain the changes in output order and serial position curves that occur with increasing list length in both ISR and IFR, thereby contributing to an integrated account of the two tasks. A key motivation for examining temporal grouping effects has been the success of Farrell's (2012) temporal clustering model, which has simulated data from a wide range of memory tasks including both IFR and ISR. Critically, to date the model is the only one to successfully simulate IFR over a wide range of list lengths, showing "ISR-like" recall in IFR of short lists and more conventional U-shaped serial position curves with longer lists. A central assumption in the model is that the majority of benchmark findings from both free recall and serial recall can be explained as a consequence of how information is structured in response to overt cues or spontaneous grouping by individuals. We next describe the model in more detail, and describe an experiment to provide a detailed comparison of grouping effects across serial and free recall tasks.

Temporal grouping in ISR and IFR

A central assumption in Farrell's (2012) model is that people spontaneously parse a sequence of information into one or more different groups of different sizes, and that the size and number of such groups determine the level of recall and the output order. Critically, and in common with suggestions in the ISR literature (Frankish, 1989; Henson, 1999; Hitch, Burgess, Towse & Culpin, 1996; Madigan, 1980; Ryan, 1969a), the grouping of items can be determined by explicit cues (e.g., temporal pauses) or spontaneously by the individual trying to remember the list. At encoding, participants associate list items with a hierarchically organised temporal context that specifies both the temporal group and the position of the item within the group. At test, it is assumed that participants access specific list items by first accessing the temporal group that is associated with the items. In an immediate test, participants are argued to have privileged access to the final group in the list, but they may also, for example, explicitly associate the context of the first group with the label "First", giving that group priority during output of the list. Exactly which group is sampled first depends upon competition between groups, but when the task is ISR, it is assumed that participants attempt to recall the first group first by "cueing" it with the "First" label.

This model is attractive in not only providing an integrated account of ISR and IFR using a hierarchical structure, but also in offering an explanation for the effects of list length and output order observed in recent work (Grenfell-Essam, Ward & Tan, 2013; Spurgeon *et al.,* 2014; Ward *et al.,* 2010). When the list is sufficiently short, it contains only one group, and recall initiates with the first item of the most recent group (or indeed the first item of the first group) leading to high tendency to initiate recall with the word in serial position 1. As the list length increases, so there is a greater need for the list to be parsed into multiple groups, such that the longer the list, generally the more groups it will contain. At recall, the

tendency to initiate recall with the first item of the most recent group will typically result in recall initiating with one of the final list items. In all cases, recall will proceed within a group in a forward serial order, as is commonly observed, even in IFR (Beaman & Morton, 2000; Bhatarah *et al.,* 2008; Farrell, 2010, 2012; Howard & Kahana, 1999; Kahana, 1996; Laming, 1999, 2006), driven by the within-group primacy mechanism.

The left hand of Figure 1 illustrates the heterogeneity of grouping structures that may spontaneously occur in ungrouped lists of different list lengths. In Figure 1, each word in a list is represented with a circle, and the group structure is indicated by the open rounded rectangles. The first word in each group is shaded grey. Farrell's (2012) model predicts that recall initiates with either the first word in the most recent group or the first word in the first list (serial position 1). As Figure 1 illustrates, for short list lengths, it is likely that there will only be one group and the first word in this first (and also most recent) group will always be serial position 1. As the list length increases, so the number of groups will increase, with the group structure at any given list length varying from trial to trial. The heterogeneous grouping structure in the ungrouped conditions results in considerable variation in the sizes of terminal groups, such that the first word in the most recent group may be any one of a number of the most recent items, leading to graded recency in the probability of first recall [P(FR)].

--Figure 1 about here --

It is not possible to know the exact grouping structure within the ungrouped lists, but the predictions that recall will tend to initiate with the first list item or the first item of the most recent group can be tested when a homogenous grouping structure is imposed upon the participants (the right hand of Figure 1). Under these conditions, the assumptions lead to strong predictions of peaks of initial recalls at the serial positions that are the first word of the

most recent group in IFR, with somewhat weaker peaks in ISR (owing to the greater tendency to use the "First" label as a cue with ISR instructions).

The core importance of temporal grouping to Farrell's (2012) model leads to the prediction that many of the characteristics of temporal grouping should be shared across different memory tasks, including ISR and IFR. Although there is considerable evidence for the effects of temporal grouping on ISR, there is very little evidence regarding the effects of temporal grouping on IFR. These studies have identified factors that influence the overall magnitude of grouping effects and the consequences of grouping on specific aspects of recall performance. They have not, however, systematically compared grouping effects in ISR and IFR across a range of equated list lengths.

In ISR, an overall serial recall advantage for grouped over ungrouped lists is regularly observed, and this effect is strongest (1) when grouping is imposed by inserting extended pauses between group boundaries (Farrell & Lewandowsky, 2004; Frankish, 1985; Henson, 1999; Hitch *et al.*, 1996; Maybery, Parmentier & Jones, 2002; Ryan, 1969a), (2) when items are grouped into regular groups of three's (Ryan, 1969b, Wickelgren, 1967), and (3) in the auditory rather than the visual modality (Frankish, 1985, 1989; Ryan, 1969a). Furthermore, spontaneous effects of grouping have been observed in ISR even with no grouping cues (e.g. Bower, 1970; Henson, 1996; Kahana & Jacobs, 2000; Madigan, 1980). One feature of grouped lists when tested by ISR is that the recall of each group resembles a 'mini-list'. Firstly, there are primacy and recency effects that occur within groups, in addition to the list as a whole, resulting in a 'scalloped' serial position curve (e.g., Frankish, 1989; Henson, 1999; Hitch *et al*, 1996; Madigan, 1980; Ryan, 1969a). Secondly, there are modality and suffix effects which occur within groups as well as to the list as a whole, with an auditory advantage for a final group item which can be abolished using a suffix (Frankish, 1985). A final feature of grouped lists when tested by ISR is that grouping reduces the overall number

of transpositions across group boundaries (Henson, 1999) and increases the number of transpositions between items that share the same within-group position (Brown *et al.,* 2000; Farrell & Lewandowsky, 2004; Henson, 1996, 1999; Johnson, 1972; Ryan, 1969a). Together, these data suggest the necessity of incorporating a multidimensional temporal structure into any complete model of working memory (Lewandowsky & Farrell, 2008), and have been key in the development of a number of models (e.g., Brown *et al.*, 2000; Burgess & Hitch, 1999, 2006; Henson, 1998), including that of Farrell (2012).

Far fewer studies have examined temporal grouping effects in IFR. In common with grouping effects in ISR, grouping effects in IFR are stronger when extended pauses are inserted between the boundaries of groups; non-temporal grouping methods such as using colour to demarcate distinct groups has a similar, but attenuated, effect (Gianutsos, 1972). However, contrary to what is commonly observed in ISR, IFR studies of temporal grouping show only a marginal overall effect. Specifically, a recall advantage for grouped items only tends to be obtained for the most recent items. This is coupled with either no effect, or even a detrimental effect, of grouping for the pre-recency items. Therefore, in IFR, there is often a non-significant effect of temporal grouping overall (Gianutsos, 1972; Tzeng & Hung, 1973).

The spontaneous (but unobservable) temporal grouping of items by participants may explain the effects of list length on serial position curves and recall order in both IFR and ISR. However, in the absence of a direct measure of the grouping pattern imposed by each person on each trial, assessing the use of spontaneous grouping on a trial-by-trial basis is challenging. Rather, our approach is to experimentally manipulate temporal grouping across a range of list lengths to assess the role of grouping in serial position and list length effects across ISR and IFR. Although the previous literature allows some suggestive preliminary comparisons, an immediate caution is that such comparisons involve a number of confounds. As with the majority of immediate memory studies, grouping experiments conducted using

ISR have typically used nine items or fewer (e.g., Frankish, 1985, 1989; Henson, 1999; Hitch *et al.,* 1996; Maybery *et al.,* 2002; Reeves, Schmauder & Morris, 2000; Ryan, 1969a, 1969b; Wickelgren, 1967), and have examined memory for closed sets of items such as digits and letters. In contrast, the grouping experiments conducted using IFR have used lists of at least twelve words (Gianutsos, 1972; Tzeng & Hung, 1973).

The aim of the current experiment was to provide a comprehensive and controlled examination of temporal grouping effects in ISR and IFR. List length was systematically manipulated from between one and 12 words. The list lengths were randomised across trials so that participants did not know the length of the list in advance of its presentation. Half of the participants performed ISR, the other half performed IFR. Within each task, half of the participants were given lists in which temporal grouping was objectively implemented by inserting an extended pause after every third word in each list, coupled with a specific instruction to think of each list as representing groups of three items. In order to maximise the effects of grouping, items were spoken as well as presented visually, and there were twice as many trials for list lengths that contained a multiple of three words (i.e., list lengths, 3, 6, 9 and 12). The other half of the participants in each task were not given such grouped lists, nor were they given grouping instructions. These ungrouped conditions served as control conditions for the respective grouped conditions, but they also allowed assessment of spontaneous grouping.

Based on the past literature, we expected to find a main effect of grouping in ISR with SR scoring with middle length lists, but far more subtle effects of grouping in IFR with FR scoring at longer lists. One particular point of interest is whether these differences in grouping effects between the tasks would be reduced when the two tasks were compared using the same list lengths and the same scoring methods. Additionally, we were interested in whether participants exhibited more subtle evidence of grouping such as mini-primacy and

mini-recency effects occurring within each group within the list (Frankish, 1989; Henson, 1999; Hitch *et al.*, 1996; Madigan, 1980; Ryan, 1969a), initiating recall of longer lists with the first item of the most recent group, and proceeding to recall within groups in forward serial order (Farrell, 2012; Henson, 1999).

Method

Participants. Eighty psychology students from the University of Essex participated in exchange for course credits.

Materials and apparatus. The materials consisted of 477 words drawn from the Toronto Noun Pool (Friendly, Franklin, Hoffman & Rubin, 1982). Subsets of 446 words were randomly selected for each participant. Using the application Supercard, items were presented visually in the centre of a Macintosh computer monitor. Simultaneous with its visual presentation, each word was presented auditorily using the digitised voice files of the Toronto Noun Pool (obtained from Kahana [2012a] at http://memory.psych.upenn.edu/WordPools). Each participant was provided with a response booklet consisting of 82 response grids, each of which contained two columns and 12 rows. The first column of each grid was narrow and contained the numbers 1-12 in ascending order. The second column was wider to allow room for participants to write down their responses.

Design. The experiment used a mixed design. There were two between-subjects independent variables: type of task with two levels (IFR and ISR) and grouping with two levels (grouped and ungrouped). There were two within-subjects independent variables: list length with twelve levels (list length 1-12), and serial position with up to twelve levels (serial position 1-12). The dependent variables were the mean number of words recalled and the proportion of words recalled using FR scoring (where a recalled word from the list was scored as correct regardless of its written position in the response grid) and the proportion of

words recalled using the relative SR scoring system used by Golomb, Peelle, Addis, Kahana & Wingfield (2008, in which a recalled word was scored as correct if it appeared later in the list than the previously recalled item, see also Drewnowski & Murdock, 1980). Note that both tasks were examined using both scoring methods, as this provides an indication of how similar performance in the two tasks is to that typically obtained in ISR (SR scoring) and IFR (FR scoring). Importantly, examining both tasks using FR scoring shows how many words were recalled in both tasks (irrespective of output position), and examining both tasks using relative SR scoring indicates the degree of forward serial order recall in both tasks. We also examined the output orders in recall. Specifically, we examined the proportions of trials that were initiated with words from each serial position and the conditionalised probabilities of transitioning at output between consecutively presented list items.

Procedure. Participants were tested individually and were informed they would be shown 2 practice lists followed by 64 experimental lists of words which they should either try to remember in the correct order (ISR) or in any order (IFR). Half of the participants performing ISR and half of the participants performing IFR were allocated to a grouped condition in which they were instructed to try to group in three's. The other half of the participants were allocated to an ungrouped condition in which they were given no such instructions. Participants were randomly allocated to conditions.

The practice lists were of 7 words. The 64 experimental lists were divided into two blocks of 32 trials. In each block, participants received two trials for each list length that was not a multiple of 3 (list lengths 1, 2, 4, 5, 7, 8, 10 and 11) and four trials for each list length that was a multiple of 3 (list lengths 3, 6, 9 and 12). Trial order within each block was randomised, so participants were not aware of the list length in advance of its presentation. The words were randomly allocated on each trial and no items were repeated across lists.

Each trial started with a warning tone and a fixation cross, followed after 1 second by a sequence of between one and 12 words simultaneously spoken by the computer and presented visually in the centre of the computer screen, during which participants remained silent. For participants in the grouped conditions, each word was presented for 0.75 s and was followed by a blank inter-stimulus interval lasting 0.25 s, except for the interval following words at serial positions 3, 6 and 9 (for relevant list lengths) which was increased to 1.25 s providing another word followed this interval. For example, the interval following the word at serial position 3 would be increased to 1.25 s for list lengths of 4 and greater as the word at serial position 4 followed the word at serial position 3; however, the length of this interval would remain at 0.25 s for list length 3 as it was the final word in the list as no word followed it.

For the ungrouped condition, the total presentation time of each list length matched that in the grouped condition. Every word was followed by an inter-stimulus interval lasting 0.25 s, but the duration that each word was presented on the screen was calculated by subtracting the sum of the inter-stimulus intervals from the total presentation time for that list length in the ungrouped condition and then dividing by the number of words in the list. Therefore each word in an ungrouped list was presented for an equal amount of time, but this time differed across different list lengths.

At the end of each list, there appeared on-screen an empty grid which resembled the grid on the response sheet but which only contained as many rows as there had been words on the list, thereby indicating at a glance the list length of the current trial. Participants wrote down as many words as they could remember on their response sheets. There was no time limit; participants finished recall when they felt like they had remembered all that they could. Participants performing IFR were free to write down their words in any temporal order that they wished and filled their response grids from the top of the grid. Participants performing

ISR were instructed to start their recall with the first item and to proceed in a forward serial order, working down the grid and writing each word in the row that corresponded to that item's serial position. If they could not remember the first item, they were asked to recall the earliest word that they could and to try to write it in the corresponding row. They were not allowed to return to fill in earlier responses following later responses.

Model simulations. As the patterns of data are more readily interpreted in the context of predictions from Farrell's (2012) model, we simulated these experiments and present the predictions along with the empirical results. Farrell (2012) presents an algorithmic description of the model, and the code for the model is included in the Supplementary Material to that paper. The model was simulated exactly as described in Farrell's (2012) simulation of the Ward *et al.* (2010) dataset (Simulation 5), with the exception that both IFR and ISR were simulated here, along with the grouping manipulation. As noted by Farrell (2012), the model does not have the facility to recall the grid positions of presented items. Accordingly, the model was simulated on a close approximation, by requiring that any items output be output in forwards order under serial recall instructions. The timing of items presented to the model was also modified in line with the experimental method. The manipulation of IFR versus ISR instructions followed that of Simulation 8 in Farrell (2012). We were interested in the qualitative predictions of the model, and so rather than fitting the model to the data obtained here, the parameter values from Simulations 5 (and 8, for manipulating task instruction) in Farrell (2012) were retained here. Accordingly, we are concerned less with the precise quantitative fit of the model, and rather focus on cases where the model either deviated substantially in its qualitative predictions, or where quantitative effects substantially differed between the model and the data. The model predictions are based on 2000 model replications.

Throughout the manuscript, we adopt a convention that empirical data are plotted on the left and the corresponding simulation of the Farrell (2012) model is plotted on the right.

Results

Mean number of words recalled. To determine whether the overall effects of temporal grouping are similar in ISR and IFR when examined over the same range of list lengths, we first focus on analyses of mean number of words recalled. Figure 2 shows the mean number of words recalled for the Grouped and Ungrouped conditions within the IFR and ISR tasks. The upper four panels represent data and simulations using free recall scoring (Figures 2A to 2D); the lower four panels represent data and simulations using relative serial recall scoring (Figures 2E to 2H). In both sets of four panels, the IFR data are presented above the ISR data.

--Figure 2 about here --

Consider first the findings of the 2 (task) x 2 (grouping) x 12 (list length) mixed ANOVA performed on the data in Figures 2A and 2C, which plot recall using FR scoring. There was no significant main effect of grouping, $F(1,76) = 2.54$, $MSE = 3.95$, $p = .115$, $\eta^2_p = .032$, or task, $F(1,76) = 3.73$, $MSE = 3.95$, $p = .057$, $\eta^2_p = .047$. The interactions between task and grouping, $F(1,76) = 0.26$, $MSE = 3.95$, $p = .609$, $\eta^2_p = .003$ and grouping and LL, $F(11,836) = 1.31$, $MSE = .426$, $p = .214$, $\eta^2_p = .009$, and the three-way interaction, $F(11, 836) = 0.72$, $MSE = .426$, $p = .719$, $\eta^2_p = .009$, were not significant. Indeed, the only significant main effect was that of list length, $F(11, 836) = 296.94$, $MSE = .426$, $p < .001$, $\eta^2_p = .796$, and the only significant interaction was between task and list length, $F(11, 836) = 1.31$, $MSE = .426$, $p = .214$, $\eta^2_p = .017$. The number of words recalled in any order increased with increasing list length in the IFR task, and increased with increasing list length in the ISR

task up to list length 5. The recall advantage in IFR was significantly greater than in ISR only at list lengths of 9 and greater.

Consider next the findings of the 2 (task) x 2 (grouping) x 12 (list length) mixed ANOVA performed on the data in Figures 2E and 2G, which plot recall using relative SR scoring. There was a significant main effect of task, $F(1, 76) = 15.44$, $MSE = 4.49$, $p < .001$, $\eta^2_p = .169$, reflecting the fact that overall more words were recalled in correct relative serial order in ISR than in IFR. There was also a significant main effect of list length, $F(11, 836) = 161.27$, $MSE = .440$, $p < .001$, $\eta^2_p = .680$, reflecting the fact that more words were recalled at longer lists than shorter lists up until list length 5, after which the number recalled in relative serial order plateaued. The interaction between task and list length was also significant, $F(11, 836) = 6.68$, $MSE = .440$, $p < .001$, $\eta^2_p = .081$, showing that the serial recall advantage in ISR was significantly greater than in IFR at list lengths of 5 and greater. The main effect of grouping, $F(1, 76) = 2.70$, $MSE = 4.49$, $p = .105$, $\eta^2_p = .034$, the interaction between task and grouping, $F(1, 76) = 0.17$, $MSE = 4.49$, $p = .684$, $\eta^2_p = .002$, and the three-way interaction, $F(11, 836) = 1.11$, $MSE = .410$, $p = .350$, $\eta^2_p = .014$ were not significant, but there was a significant interaction between grouping and list length, $F(11, 836) = 1.91$, $MSE = .440$, $p = .035$, $\eta^2_p = .025$, reflecting a significant SR advantage for the grouped lists over the ungrouped list at list lengths 5, 6 and 12.

The plots of the model simulations (right hand panels) show that the model over-predicts the effects of grouping using both the FR and the relative SR scoring. More problematic (and informative) is that the model predicts that the number of words recalled in IFR increases as a function of the number of items in the most recent cluster. This saw-toothed pattern is driven by trials in which participants initiate recall with a word from the most recent cluster, and occurs because output interference in Farrell's (2012) model occurs at the level of temporal groups rather than individual items: participants can output the

contents of the current cluster (whether that be 1, 2, or 3 words) with the same degree of output interference[1]. Although the data do not rule out output interference at the level of groups, they do imply that the majority of output interference occurs at the level of individual items.

The probability of first recall (P[FR]). Analyses of the P(FR) data are important for determining where participants initiate their recall. A key prediction from the Farrell (2012) model is that participants initiate their recall with the first word of the most recent group, in violation of the "standard" recency effect. Tables 1 and 2 show the proportion of trials in which words from different serial positions were recalled first for each list length for the grouped and ungrouped conditions for the ISR and IFR tasks, respectively.

--Tables 1 and 2 about here--

The P(FR) data from Tables 1 and 2 were collapsed into one of three categories: 'SP1' (recall started with the first word in the list), 'Last 4' (recall started with one of the last four items in the list; note that for LLs 2-4 this included all of the items except for the first word in the list), and 'Other' (recall started with any of the other list items, or began with an intrusion, or nothing was recalled). Figure 3 shows the proportion of trials in which words from different list positions were recalled first as a function of list length for each of these three categories. The upper panels show the data and simulations of the IFR task, the lower panels show the data and simulations of the ISR task.

--Figure 3 about here--

As can be seen, in all four sets of data (left hand plots), participants show a strong

[1] Data and simulations most relevant to this specific explanation are outlined in the Supplementary materials.

tendency to initiate their recall with the first list item (serial position 1) for short to medium list lengths. This tendency decreases with increasing list length in all four conditions, but decreases more strongly in the IFR conditions, such that there is a cross-over in the PFR data for IFR: the list length at which the modal response changed from 'SP1' to 'Last 4' occurred at around list length 6 for the grouped condition and list length 7 for the ungrouped condition. Although the tendency to initiate recall with the first list item also declines with increasing list length in the ISR conditions, there was no cross-over in either of the ISR conditions demonstrating that, even at the longer list lengths, when participants are given ISR instructions they comply by starting with the first item on the majority of trials.

Considering the 2 (task: IFR and ISR) x 2 (grouping: grouped and ungrouped) x 12 (list length: 1-12) mixed ANOVA based on the proportion of responses where recall was initiated with the first list item (i.e. P[FR=SP1]), the proportion of trials in which P(FR=SP1), the main effects revealed the tendency to initiate recall at the start of the list was greater in ISR relative to IFR, $F(1, 76) = 47.21$, $MSE = .291$, $p < .001$, $\eta^2_p = .383$, and greater at short list lengths, $F(11, 836) = 142.91$ $MSE = .037$, $p < .001$, $\eta^2_p = .653$. The significant interaction between task and list length revealed that that the tendency was greater in ISR relative to IFR for all list lengths greater than 3, $F(11, 836) = 13.95$, $MSE = .037$, $p < .001$, $\eta^2_p = .155$. The main effect of grouping, $F(1,76) = 1.47$, $MSE = .291$, $p = .229$, $\eta^2_p = .019$, the interactions between task and grouping $F(1,76) = 0.78$, $MSE = .291$, $p = .380$, $\eta^2_p = .010$, grouping and list length, $F(11, 836) = 0.52$, $MSE = .037$, $p = .894$, $\eta^2_p = .007$, and the three-way interaction, $F(11, 836) = 1.55$, $MSE = .037$, $p = .109$, $\eta^2_p = .020$, failed to reach significance, demonstrating that the grouping manipulation did not affect participants' tendency to initiate recall with the first list item. The equivalent analyses for the proportion of trials in which P(FR=Last 4) can be found in the supplementary materials, but to summarise, the effects were complimentary to the P(FR=SP1) in that the tendency to initiate recall with one of the

last four items was greater in IFR relative to ISR and greater at longer list lengths. Specifically, the tendency was greater in IFR relative to ISR for all list lengths greater than 3. None of the main effects or interactions involving grouping were significant, demonstrating that the grouping manipulation did not affect participants' tendency to initiate recall with one of the last four list items.

Farrell's (2012) model is in broad agreement with these coarse-grained P(FR) data, especially for IFR. Although the cross-over point is predicted to be somewhat lower than the data (between list lengths 3 and 4 in the simulations), the Farrell model correctly predicts that at longer list lengths, participants will initiate recall with one of the last 4 words in IFR at longer list lengths. The Farrell (2012) model also correctly predicts the absence of a cross-over in the P(FR) in the ISR data (bottom half of figure). However, the model over-predicts the magnitude of the grouping manipulation. When the list length exceeds a multiple of 3 (e.g., moving from list length 3 to list length 4), the model predicts a substantial drop in the frequency of 'SP1' categories: these are cases where a new group is added to the end of the list, and this group is given substantial priority (being the most recent group) over the first group in the list. The model also severely under-predicts the proportion of trials in which recall starts with a word from an "Other" serial position.

A more detailed illustration of the fine-grained P(FR) data for the IFR task is shown in Figure 4, which plots the data as a function of serial position. Again, the left-hand panels show the data and the right-hand panels show the corresponding simulations; the upper panels show the data and simulations of the IFR task, the lower panels show the data and simulations of the ISR task. What is very apparent in Figure 4 is that there are noticeable peaks at serial positions 4, 7 and 10 in the grouped conditions, particularly in the IFR data at long lists. This demonstrates an increased tendency to initiate recall with the first item from the most recent group. This is a key prediction from the Farrell (2012) model and it is clearly

apparent from the simulation. [2]

--Figure 4 about here--

If anything, the model over-estimates the tendency to initiate with the first word in the most recent cluster (especially in the ISR data). A close examination of Tables 1 and 2, show that in the data from the grouped conditions, there are indeed increased tendencies to initiate recall with the first item from a cluster, but this tendency extends beyond the most recent cluster to additionally include earlier groups (as indicated by the increase in the underlined values in the grouped conditions). It would appear that participants do tend to initiate recall with the first word in a temporal cluster, but that there is better than predicted initial accessibility to clusters other than the first or last. A further difference in the IFR data is that at list lengths 4, 5 and 6, despite a peak appearing at serial position 4, participants are still most likely to initiate their recall with the first list item; in contrast, the model shows a substantial drop in recall of the first list item beyond list length 3, mirroring the same drop in the aggregate analysis plotted in Figure 3.

Analyses of serial position curves. Analyses of the serial position curves provide comparisons of the magnitude and similarity of the primacy and recency effects between grouped and ungrouped conditions in ISR and IFR, and the extent to which these are modulated by grouping. Figure 5 shows the serial position curves for all list lengths for the grouped and ungrouped conditions in IFR and ISR using FR scoring. The left-hand panels show the data and the right-hand panels show the corresponding simulations; the upper panels show the data and simulations of the IFR task, the lower panels show the data and simulations of the ISR task. Within each task, the grouped and ungrouped conditions are

[2] A series of 2 (task: IFR and ISR) x 2 (grouping: grouped and ungrouped) x *n* (serial position: 1-*n*) mixed ANOVAs, where *n* is the list length on the P(FR) data at each list length can be found in supplementary materials.

plotted separately. The corresponding serial position curves of the data and simulations plotted using relative SR scoring are shown in Figure 6.

--Figures 5 and 6 about here--

Considering first the data with FR scoring (Figure 5), the serial position curves for the data (left hand panels) changed in similar ways with increasing list length in all four conditions. Performance was at ceiling for the very short list lengths (i.e. list lengths 1-4), but as list length increased there were primacy and recency effects at shorter list lengths; and then reduced primacy effects and increased recency effects at longer list lengths. There were greater primacy effects in ISR relative to IFR, and greater recency effects in IFR relative to ISR. The effects of grouping in both tasks were hard to detect using FR scoring.

Considering next the data using relative SR scoring (Figure 6), the serial position curves again changed in similar ways with increasing list length. For ISR, there were primacy effects with reduced recency effects at shorter list lengths, and then reduced primacy effects and increased recency effects at longer list lengths. For IFR, there were recency effects with limited primacy effects at short list lengths, and then virtually no primacy effects with increased recency effects at longer list lengths. Again, the effects of grouping were minimal.[3]

The simulations of the ungrouped data reasonably approximate the data, with the exception that the model underestimates recency in the ISR task. More problematic is that the model over-exaggerates the discontinuities at the group boundaries in the grouped data. When one looks carefully at the data, one can see hints of discontinuities at the group boundaries in the group conditions (particularly at longer list lengths), but the simulation dramatically over-emphasises these scalloping effects.

[3] A series of 2 (task: IFR and ISR) x 2 (grouping: grouped and ungrouped) x *n* (serial position: 1-*n*) mixed ANOVAs, where *n* is the list length for both FR and relative SR scoring at each list length can be found in supplementary materials, for both all data, and for just the proportion of trials in which P(FR=SP1) and P(FR=Last 4).

Conditionalised response probabilities (CRPs). Previous studies have shown that recall of an item *N* tends to be followed by recall of an item presented at a nearby serial position, particularly the following one (*N*+1). Farrell's (2012) model produces this forward recall tendency by virtue of the forward recall of items within groups. A key prediction of the model is that there will be more within-group transitions and fewer across-group transitions in the grouped relative to the ungrouped conditions, as items within a group will tend to be recalled together.

We examined the extent to which the output order showed evidence of forward serial order recall and transitions between different serial positions which maintained within-group position, by calculating CRP values at different lags (for a more detailed description of lag-CRP analyses, see Howard & Kahana, 1999; Kahana, 1996; Kahana, Howard & Polyn, 2008). The lag refers to the difference in serial position based on the input positions of the words that were recalled (i.e., the difference between successive pairs of words, which is calculated by subtracting the serial position of the first word of each pair of responses from the serial position of the second word of each pair). Smaller lag values therefore represent recall transitions between words from closer positions, whereas larger lag values represent recall transitions between words from more distant positions. In addition, positive lag values represent recall transitions proceeding in a forward direction; negative lag values represent recall transitions proceeding in a backward direction.

In order to calculate the CRP at each lag, for every correct response, both the participants' actual transitions and possible legitimate transitions at a given lag were recorded. The actual number of specific lag transitions was divided by the number of opportunities to output that specific transition. An opportunity to output a specific lag transition was dependent whether the lag transition was actually possible (i.e. within the valid

range of serial positions). This procedure controls for the increased opportunities to make transitions at small lags and the reduced opportunities to make transitions at extreme lags.

We present here the most diagnostic analyses, in which we consider the CRP of making Lag +1 responses when recall continues within a group (i.e., a transition between serial position 2 and 3, 5 and 6 or 8 and 9) or across a group boundary (i.e., a transition between serial position 3 and 4, 6 and 7, or 9 and 10). If grouping drives dependencies between successive recall attempts, we should expect to see a fall-off in the CRP when transiting from the final item in a group, as there is no special tendency to recall items from the following group. Figure 7 shows the proportion of CRP Lag+1 transitions within and across group boundaries in the grouped and ungrouped conditions for list lengths of 5 and greater. The left-hand panels show the data and the right-hand panels show the corresponding simulations; the upper panels show the data and simulations of the IFR task, the lower panels show the data and simulations of the ISR task.

--Figure 7 about here--

Two separate 2 (grouping: grouped and ungrouped) x 2 (transition type: within-group and across-group) x 8 (list length: 5-12) mixed ANOVA were performed on the proportion of Lag + 1 transitions for first the IFR and then the ISR data. For IFR, the main effects revealed there were more within-group +1 transitions relative to across-group +1 transitions, $F(1,38) = 11.63$, $MSE = .100$, $p = .002$, $\eta^2_p = .234$, and the main effect of list length again revealed a somewhat zig-zag pattern, $F(7, 266) = 7.88$, $MSE = .114$, $p < .001$, $\eta^2_p = .172$. The main effect of grouping failed to reach significance, $F(1,38) = 1.09$, $MSE = .264$, $p = .304$, $\eta^2_p = .028$. The significant interaction between grouping and transition type, $F(1,38) = 20.94$, $MSE = .100$, $p < .001$, $\eta^2_p = .355$, revealed that the greater number of within-group transitions relative to across-group transitions occurred in grouped lists only; the effect of transition type

was not significant for ungrouped lists. The significant interaction between transition type and list length, $F(7, 266) = 4.24$, $MSE = .093$, $p < .001$, $\eta^2_p = .100$, revealed that the greater number of within-group transitions relative to across group transitions occurred at list lengths 6, 9 and 12 only. The interaction between grouping and list length $F(7, 266) = 1.57$, $MSE = .114$, $p = .144$, $\eta^2_p = .040$, and the three-way interaction, $F(7, 266) = 1.35$, $MSE = .093$, $p = .226$, $\eta^2_p = .034$, failed to reach significance.

For ISR, the main effects revealed there were more within-group +1 transitions relative to across-group +1 transitions $F(1,38) = 6.38$, $MSE = .089$, $p = .016$, $\eta^2_p = .144$, and the main effect of list length, $F(7, 266) = 14.68$, $MSE = .074$, $p < .001$, $\eta^2_p = .279$, revealed a somewhat zig-zag pattern, with an overall tendency for a drop in the functions with increasing list length. The main effect of grouping failed to reach significance, $F(1,38) = 3.99$, $MSE = .308$, $p = .053$, $\eta^2_p = .095$. The significant interaction between transition type and list length, $F(7, 266) = 1.65$, $MSE = .062$, $p < .001$, $\eta^2_p = .112$, revealed that the greater number of within-group transitions relative to across group transitions occurred at list lengths 6 and 9. The interaction between grouping and transition type, $F(1,38) = 2.17$, $MSE = .089$, $p = .149$, $\eta^2_p = .054$, and grouping and list length, $F(7, 266) = 1.65$, $MSE = .074$, $p = .122$, $\eta^2_p = .042$, and the three-way interaction, $F(7, 266) = 0.41$, $MSE = .062$, $p = .899$, $\eta^2_p = .011$, failed to reach significance.

The model's predictions are broadly in accord with the data. In IFR, the model predicts a difference between within-group and across-group +1 transitions only for the grouped condition. The model also predicts the negligible interaction involving grouping in ISR, and the overall higher frequency of within-group transitions. However, the model over-predicts the frequency of +1 transitions generally in ISR, and predicts a main effect of grouping that is less evident (only marginally significant) in the data.

General Discussion

This research investigated the effects of temporal grouping on the accuracy and the output order of recall in ISR and IFR over the same range of list lengths. Recent work has suggested that theoretical accounts of STM and working memory - including the seminal work of Baddeley and Hitch (1974) and other broad theories (e.g., Cowan, 1999) - would benefit from integrating both ISR and IFR. One such model is that of Farrell (2012), which highlights that temporal grouping processes play a prominent role in the patterns of recall and output orders in ISR and IFR. This discussion will first focus on the grouping effects observed in the IFR and ISR experimental data. We then discuss the theoretical implications of the successes and failures of the Farrell (2012) in accounting for our data.

Temporal grouping effects in ISR and IFR

Consistent with previous studies of temporal grouping on ISR (e.g. Frankish, 1985, 1989; Henson, 1999; Hitch *et al.*, 1996; Maybery *et al.*, 2002; Reeves *et al.*, 2000; Ryan, 1969a, 1969b; Wickelgren, 1967), an overall advantage of grouping in ISR was found with relative SR scoring, most clearly at medium list lengths typically used in ISR. However, the grouping effect was somewhat weaker than is generally observed in ISR grouping experiments, possibly because the open set size entailed more reliance on item information, whereas studies using a closed set of stimuli only involve the maintenance of order information, which is where grouping effects are most apparent (e.g. Frankish, 1985, 1989; Henson, 1999; Hitch *et al.*, 1996; Maybery *et al.*, 2002; Parmentier & Maybery, 2008; Ryan, 1969a, 1969b; Wickelgren, 1967).

Consistent with previous studies of temporal grouping on IFR (e.g., Gianutsos, 1972; Tzeng & Hung, 1973), we found no overall advantage of grouping in IFR with FR scoring. When our ISR data was similarly analysed using FR scoring, there was also no effect of

grouping with FR scoring, in contrast to the effect observed in ISR when using relative SR scoring. This suggests that one reason for the different effects of grouping in the IFR and ISR literatures was of the differences in the scoring systems that are typically used. These earlier IFR studies had shown a grouping advantage at recency positions coupled with a detrimental or non-significant effect at primacy positions. Broadly consistent with these earlier findings, we found that participants who were presented with longer lists for a test of IFR showed a strong tendency to initiate recall with the first word from the most recent group, followed by a forward run of recency items within the group. However, these effects of grouping on the overall serial position curves were somewhat diluted in our data, relative to earlier studies, possibly because our lists were shorter in length than those used by Gianutsos (1972) and Tzeng and Hung (1973), and for shorter lists there was a larger proportion of trials in which recall started with the first list item, resulting in only a subset of trials which showed the previously reported IFR grouping effect.

In our analyses, we found no significant interactions involving task and group: we found no effects of grouping with FR scoring, and we also found effects of grouping only with relative SR scoring at middle list lengths (i.e., those typically used in grouping studies of ISR). These findings suggest that the differences in the scoring and differences in the list lengths may have contributed to the recorded differences in the literature of grouping effects with SR scoring at middle list lengths, and no grouping effects with FR scoring at longer list lengths.

Consistent with recent work (Grenfell-Essam & Ward, 2013; Spurgeon *et al*., 2014; Ward *et al*., 2010), there were many clear similarities between our ISR and IFR data. First, there were strong effects of list length on the accuracy and output order of recall in both tasks. With shorter lists on both tasks, there was an increased tendency to initiate recall with the first word in the list and proceed in forward order. The tendency to initiate recall with the

first list item was far stronger in ISR than in IFR, reflecting that participants on the whole could carry out ISR instructions even at longer list lengths. As the list length was increased, so this tendency weakened in both tasks, and in IFR the modal tendency on longer lists was to initiate recall with one of the last four words, and when this occurred there was an enhanced tendency to recall other recent items. In addition, in the data for both tasks, the grouping manipulation did not affect the probability of initiating recall with the first list item, but did increase the probability of making transitions within groups and a reduction in the probability of transitions across group boundaries.

Relations to the Farrell (2012) model, and theoretical implications

The Farrell (2012) model reasonably accounted for many of the main features of both IFR and ISR. The model was generally accurate at predicting the list length effects and the tendency to initiate recall at different list lengths, mirroring the overall concordance between model and data seen in simulations of similar data sets (see the simulation of the Ward *et al.*, 2010, data set in Farrell, 2012). In particular, the model correctly predicts a peak in the P(FR) for items at the beginning of each group in IFR (Figure 4), consistent with the assumption that people attempt to serially recall the items from each group. At the shorter list lengths, this peak occurred at the first serial position demonstrating the strong tendency to initiate recall with the first list item, whereas at longer list lengths, this peak tended to occur at the serial position of the first item of the most recent group. The model was also reasonably accurate at predicting the decreased tendency to transit between successive items that straddled a group boundary, another key signature of the important role of the grouping structure in determining patterns of recall.

However, comparison of the model and the data suggest that the model's performance was more sensitive to the grouping manipulation than was the behaviour of our participants.

People initiated recall with the first list item more often than predicted by the model, although this may reflect strategic differences that are easily captured by varying ω, a parameter in the model capturing people's preference for beginning recall with the last group. A more interesting departure from the data is the pronounced drop in the P(FR) at serial position 1 (Figure 3) when the list length exceeds 3 or 6 (multiples of the group size), which contrasts with the data that show a more gradual drop in this point as list length increases. The over-emphasis of temporal grouping in the model was also seen in the serial position functions; in contrast to the data, which showed relatively smooth functions, the model produced a more scalloped pattern, including a relatively flat function across the items in the last group.

One more "productive failure" of the model was its prediction that the capacity limitation introduced by output interference operates at the level of groups, and not individual items. This produces the zig-zag pattern seen in Figure 2 that is not readily apparent in the data. The assumption of group-level output interference is required to explain apparent group-level effects in other data sets (e.g., Cowan, Saults, Elliott, & Moreno, 2002), but the present data suggest a large degree of output interference in free recall occurs at the level of individual items. What is not clear, and must remain a topic for further research, is whether the relative effects of group-level and item-level output interference relate to the nature of the memoranda (words vs. letters/digits, open vs. closed sets), the task used (serial recall vs. free recall), or some other factor.

Together, the data support the key assumption of the model - the central role of grouping in determining recall - and are consistent with the idea that the principles of hierarchical organization central to theorising in short-term memory (e.g., Brown *et al.*, 2000; Burgess & Hitch, 1999; Henson, 1998) also apply to recall of longer sequences, and irrespective of whether serial recall is required. The Farrell (2012) model provides 'existence proof' that an integrated account of ISR and IFR, and other STM memory tasks, can explain

the data. However, the data also place important constraints on the model. The sensitivity of the Farrell model to grouping effects follows from its ability to account for fine-grained aspects of serial recall data - including the numerous effects of grouping outlined in the introduction - whilst also explaining free recall data. Although weakening the effects of grouping in the model might lead to a better fit here, this would be at a cost to accounting for grouping effects in "typical" ISR in the same invariant model. The two alternative implications are that a) the mechanisms driving ISR and IFR differ in some fashion, or b) other differences between the tasks produce evidence for (or lack thereof) the effects of temporal grouping. Along with other issues left open by the current study - for example, which characteristics of sequences play a more critical role in determining how participants parse the elements into clusters - this central issue must be answered by future research.

References

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

Atkinson, R. C., & Shiffrin, R. M. (1971). The control of short-term memory. *Scientific American, 225,* 82-90.

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

Baddeley, A. D. (1986). *Working Memory*. Oxford: Clarendon Press.

Baddeley, A. (2012). Working memory: Theories, models and controversies. *Annual Review of Psychology, 63,* 1-29.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. A. Bower (Ed.) *Recent advances in learning and motivation, Vol. 8.* (pp. 47-90). London: Academic Press.

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior, 14,* 575-589.

Beaman, C. P., & Morton, J. (2000). The separate but related origins of the recency and the modality effect in free recall. *Cognition*, 77, B59-B65.

Bhatarah, P., Ward, G., Smith, J., & Hayes, L. (2009). Examining the relationship between free recall and immediate serial recall: Similar patterns of rehearsal and similar effects of word length, presentation rate, and articulatory suppression. *Memory and Cognition, 37,* 689-713.

Bhatarah, P., Ward, G., & Tan, L. (2008). Examining the relationship between free recall and immediate serial recall: The serial nature of recall and the effect of test expectancy. *Memory & Cognition, 36,* 20-34.

Botvinick, M. M., & Plaut, D. C. (2006). Short-term memory for serial order: A recurrent neural network model. *Psychological Review, 113,* 201-233.

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

Broadbent, D. E. (1975). The magic number seven after fifteen years. In A. Kennedy & A. Wilkes (Eds.), *Studies in long-term memory* (pp. 3-18). Oxford: John Wiley & Sons.

Brown, G. D. A., Chater, N., & Neath, I. (2008). Serial and free recall: Common effects and common mechanisms? A reply to Murdock (2008). *Psychological Review, 115,* 781-785.

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

Brown, G. D. A., Preece, T., & Hulme, C. (2000). Oscillator-based memory for serial order. *Psychological Review, 107,* 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,* 551-581.

Burgess, N., & Hitch, G. J. (2006). A revised model of short-term memory and long-term learning of verbal sequences. *Journal of Memory and Language, 55,* 627-652.

Chen, Z., & Cowan, N. (2009). How verbal memory loads consume attention. *Memory & Cognition, 37,* 829-836.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information processing system. *Psychological Bulletin, 104,* 163-191.

Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.) *Models of Working Memory: Mechanisms of active maintenance and executive control* (pp. 62-101). Cambridge, U.K.: Cambridge University Press.

Cowan, N. (2000). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24,* 87-185.

Cowan, N. (2005). *Working memory capacity.* Hove, East Sussex, UK: Psychology Press.

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

Crannell, C. W., & Parrish, J. M. (1957). A comparison of immediate memory span for digits, letters, and words. *Journal of Psychology, 44,* 319-327.

Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H. J., & Usher, M. (2005). The demise of short-term memory revisited: Empirical and computational investigations of recency effects. *Psychological Review, 112,* 3–42.

Drewnowski, A., & Murdock, B. B., Jr. (1980). The role of auditory features in memory span for words. *Journal of Experimental Psychology: Human Learning and Memory, 6,* 319–332.

Farrell, S. (2010). Dissociating conditional recency in immediate and delayed free recall: A challenge for unitary models of recency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 324-347.

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

Farrell, S., & Lewandowsky, S. (2002). An endogenous distributed model of ordering in serial recall. *Psychonomic Bulletin & Review, 9,* 59-79.

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

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

Frankish, C. (1989). Perceptual organization and precategorical acoustic storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 469-479.

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

Gianutsos, R. (1972). Free recall of grouped words. *Journal of Experimental Psychology, 95,* 419-428.

Glanzer, M. (1972). Storage mechanisms in recall. In G. H. Bower, (Ed.) *The psychology of learning and motivation: Advances in research and theory*. (Vol. 5. pp. 129-193). New York: Academic Press.

Golomb, J. D., Peelle, J. E., Addis, K. M., Kahana, M. J., & Wingfield, A. (2008). Effects of adult aging on utilization of temporal and semantic associations during free and serial recall. *Memory & Cognition, 36,* 947–956.

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.

Grenfell-Essam, R., Ward, G., & Tan, L. (2013). The role of rehearsal on the output order of immediate free recall of short and long lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39,* 317-347.

Grossberg, S., & Pearson, L. R. (2008). Laminar cortical dynamics of cognitive and motor working memory, sequence learning and performance: Toward a unified theory of how the cerebral cortex works. *Psychological Review, 115,* 677-732.

Henson, R. N. A. (1996). Short-term memory for serial order (Unpublished doctoral dissertation). University of Cambridge, Cambridge, England.

Henson, R. N. A. (1998). Short-term memory for serial order: The start-end model of serial recall. *Cognitive Psychology, 36,* 73-137.

Henson, R. N. (1999). Positional information in short-term memory: relative or absolute? *Memory & Cognition, 27,* 915-927.

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,* 116-139.

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 923-941.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46,* 269-299.

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,* 339-373.

Johnson, N. F. (1972). Organization and the concept of a memory code. In A. W. Melton and E. Martin (Eds.), *Coding processes in human memory* (pp. 125-159). Washington, DC: Winston.

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition, 24,* 103-109.

Kahana, M. J. (2012a). Auditory Toronto word pool. Available : http://memory.psych.upenn.edu/WordPools (11th September, 2012)

Kahana, M. J. (2012b). *Foundations of Human Memory.* New York: Oxford University Press.

Kahana, M. J., Howard, M. J., & Polyn, S. M. (2008). Associative retrieval processes in episodic memory. In J. Byrne (Series Ed.) *Learning and memory: A comprehensive*

reference. Vol 2: Cognitive psychology of memory (H. L. Roediger, III, Vol. Ed.). Oxford: England: Elsevier.

Kahana, M. J., & Jacobs, J. (2000). Interresponse times in serial recall: Effects of intraserial repetition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1188–1197.

Laming, D. (1999). Testing the idea of distinct storage mechanisms in memory. *International Journal of Psychology, 34,* 419-426.

Laming, D. (2006). Predicting free recalls. *Journal of Experimental Psychology: Learning, Memory and Cognition, 32,* 1146-1163.

Lewandowsky, S., & Farrell, S. (2008). Short-term memory: New data and a model. *The Psychology of Learning and Motivation, 49,* 1-48.

Madigan, S. A. (1980). The serial position curve in immediate serial recall. *Bulletin of the Psychonomic Society, 15,* 335-338.

Maybery, M. T., Parmentier, F. B. R., & Jones, D. M. (2002). Grouping of list items reflected in the timing of recall: Implications for models of serial verbal memory. *Journal of Memory and Language, 47,* 360-385.

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

Murdock, B. B., Jr. (1962). The serial position effect of free recall. *Journal of Experimental Psychology, 64,* 482-488.

Page, M. P. A., & Norris, D. (1998). The primacy model: A new model of immediate serial recall. *Psychological Review, 105,* 761-781.

Page, M. P. A., & Norris, D. G. (2003). The irrelevant sound effect: What needs modeling, and a tentative model. *The Quarterly Journal of Experimental Psychology, 56A,* 1289-1300.

Parmentier, F. B. R., & Maybery, M. T. (2008). Equivalent effects of grouping by time, voice and location on response timing in verbal serial memory. *Journal of Experimental Psychology: Learning, Memory & Cognition, 34,* 1349-1355.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88,* 93–134.

Reeves, C., Schmauder, A. R., & Morris, R. K. (2000). Stress grouping improves performance on an immediate serial list recall task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1638-1654.

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology, 89,* 63-77.

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

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

Spurgeon, J., Ward, G., & Matthews, W. J. (2014). Examining the relationship between immediate serial recall and immediate free recall: Common effects of phonological loop variables but only limited evidence for the phonological loop. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 1110-1141.

Tan, L., & Ward, G. (2000). A recency-based account of primacy effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1589-1625.

Tzeng, O. J. L., & Hung, D. I. (1973). Intralist organization and subsequent free recalls. *Journal of Experimental Psychology, 98,* 119-124.

Ward, G., Tan, L., & Grenfell-Essam, R. (2010). Examining the relationship between free recall and immediate serial recall: The effects of list length and output order. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 1207-1241.

Waugh, N. C., & Norman, D. A. (1965). Primary memory. *Psychological Review, 72,* 89-104.

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

Tables

Table 1. The distribution of the first words recalled on each trial for the grouped and ungrouped conditions for the ISR task as a function of the list length.

Serial position	List length											
	1	2	3	4	5	6	7	8	9	10	11	12
						ISR Grouped						
1	**80**	**80**	**78**	***77***	***76***	***68.5***	*58*	*55*	*49*	*43*	*42*	*37.5*
2		0	2	3	3	2.5	1	3	4	2	4	5
3			0	0	0	3	3	1	2	5	1	0.5
4				**0**	**0**	**3**	***12***	***10***	***6.5***	*6*	*6*	*8.5*
5					0	0.5	2	3	3.5	3	3	2
6						0.5	0	4	3	1	4	2
7							**0**	**1**	**6.5**	***13***	***10***	***9***
8								0	0	2	2	0
9									2	1	1	4
10										**0**	**3**	**6**
11											1	1
12												0.5
Void/error	0	0	0	0	1	2	4	3	3.5	4	3	4
Total	80	80	80	80	80	80	80	80	80	80	80	80
						ISR Ungrouped						
1	**80**	**78.5**	**78.5**	***71***	***73***	***66***	*58*	*51*	*58*	*51*	*44*	*45.5*
2		1	0.5	6	1	3	5	3	3	2	5	3.5
3			0.5	0	5	3.5	5	4	3	1	1	3
4				**0**	**1**	**3**	***3***	***6***	***1***	*2*	*1*	*0.5*
5					0	2	2	6	4	3	5	2
6						1	2	5	4	7		1.5
7							**1**	**1**	**2**	***6***	***4***	***5***
8								2	1.5	4	3	2.5
9									0.5	2	7	4.5
10										**0**	**2**	**4**
11											2	3
12												1
Void/error	0	0.5	0.5	3	0	1.5	4	2	3	2	6	4
Total	80	80	80	80	80	80	80	80	80	80	80	80

Note, the **bold** values represent the frequency of trials in which the first word recalled was the first word from the most recent group, the ***bold underlined italicised*** values represent the frequency of trials in which the first word recalled was the first word from the penultimate group, the *underlined italicised* values represent the frequency of trials in which the first word recalled was the first word from the pre-penultimate group, and the *italicised* values represent the frequency of trials in which the first word recalled was the first word from the group before the pre-penultimate group. Void = no words were recalled on a particular trial; error = word recalled not on the list.

Table 2. The distribution of the first words recalled on each trial for the grouped and ungrouped conditions for the IFR task as a function of the list length.

Serial position	List length											
	1	2	3	4	5	6	7	8	9	10	11	12
						IFR Grouped						
1	**80**	**79**	**76**	***61***	***48***	***38***	*27*	*21*	*16*	*15*	*12*	*11*
2		1	2.5	4	3	5	3	2	2	2	3	1.5
3			1.5	1	4	4	4	3	2	1	2	0.5
4				**13**	**21**	**20**	***13***	***15***	***5***	*3*	*3*	*5*
5					3	5.5	2	0	1	1	1	0
6						5.5	4	4	2	1	0	0
7							**27**	**30**	**38**	***16***	***10***	***6.5***
8								4	5.5	1	2	1.5
9									6.5	4	2	2
10										**35**	**43**	**38.5**
11											2	5
12												6.5
Void/error	0	0	0	1	1	2	0	1	2	1	0	2
Total	80	80	80	80	80	80	80	80	80	80	80	80
						IFR Ungrouped						
1	**80**	**80**	**77**	***69***	***62***	***49.5***	*39*	*30*	*21.5*	*16*	*18*	*13*
2			1.5	3	4	3	2	4	2	1	1	1.5
3			0.5	3	4	2.5	3	3	2.5	2	2	1
4				**5**	**6**	**5**	***6***	***2***	***3.5***	*3*	*1*	*3.5*
5					3	9	8	5	3	3	1	0
6						10	6	8	6.5	4	4	2.5
7							**15**	**11**	**9**	***7***	***4***	***4***
8								15	6	9	6	6.5
9									23.5	12	9	6.5
10										**22**	**9**	**3.5**
11											24	13
12												24.5
Void/error	0	0	1	0	1	1	1	2	2.5	1	1	0.5
Total	80	80	80	80	80	80	80	80	80	80	80	80

Note, the **bold** values represent the frequency of trials in which the first word recalled was the first word from the most recent group, the ***bold underlined italicised*** values represent the frequency of trials in which the first word recalled was the first word from the penultimate group, the *underlined italicised* values represent the frequency of trials in which the first word recalled was the first word from the pre-penultimate group, and the *italicised* values represent the frequency of trials in which the first word recalled was the first word from the group before the pre-penultimate group. Void = no words were recalled on a particular trial; error = word recalled not on the list.

Figure Captions

Figure 1. An illustration of the heterogeneity of group structures at different list lengths in the Ungrouped Lists (Left hand side) and the homogeneity of group structures at different list lengths in the Grouped Lists (Right hand side). Individual words are illustrated by circles, groups are illustrated by open rounded rectangles. In all lists, participants are assumed to initiate recall with the first word of a group (illustrated by grey circles); most often the first word of the first group, or the first word in the current / most recent group.

Figure 2. The mean number of words recalled from lists of one to 12 words from the grouped and ungrouped conditions. The left-hand panels show the data from IFR task (FR scoring: Figure 2A, relative SR scoring: Figure 2E) and the ISR task (FR scoring: Figure 2C, relative SR scoring: Figure 2G); and the right-hand panels show the Farrell (2012) model simulations from the IFR task (FR scoring: Figure 2B, relative SR scoring: Figure 2F) and the ISR task (FR scoring: Figure 2D, relative SR scoring: Figure 2H).

Figure 3. The P(FR) data showing the proportion of trials in which recall was initiated with the first word in the list, one of the last four words in the list, or any other item. Each condition is plotted separately as a function of list length. The left-hand panes are the data from the IFR ungrouped condition (Figure 3A), the IFR grouped condition (Figure 3C), the ISR ungrouped condition (Figure 3E) and the ISR grouped condition (Figure 3G); and the right-hand panels are the model simulations from the IFR ungrouped condition (Figure 3B), the IFR grouped condition (Figure 3D), the ISR ungrouped condition (Figure 3F) and the ISR grouped condition (Figure 3H).

Figure 4. The P(FR) data showing the proportion of trials in which recall was initiated with each serial position. Each condition is plotted separately as a function of list length. The left-hand panels are the data from the IFR ungrouped condition (Figure 4A), the IFR grouped condition (Figure 4C), the ISR ungrouped condition (Figure 4E) and the ISR grouped condition (Figure 4G); and the right-hand panels are the model simulations from the IFR ungrouped condition (Figure 4B), the IFR grouped condition (Figure 4D), the ISR ungrouped condition (Figure 4F) and the ISR grouped condition (Figure 4H).

Figure 5. The serial position curves from lists of one to 12 words in the grouped and ungrouped conditions of the ISR and IFR tasks. The eight panels show the serial position curves using FR scoring. The left-hand panels are the data from the IFR ungrouped condition (Figure 5A), the IFR grouped condition (Figure 5C), the ISR ungrouped condition (Figure 5E) and the ISR grouped condition (Figure 5G); and the right-hand panels are the model simulations from the IFR ungrouped condition (Figure 5B), the IFR grouped condition (Figure 5D), the ISR ungrouped condition (Figure 5F) and the ISR grouped condition (Figure 5H).

Figure 6. The serial position curves from lists of one to 12 words in the grouped and ungrouped conditions of the ISR and IFR tasks. The eight panels show the serial position curves using relative SR scoring. The left-hand panels are the data from the IFR ungrouped condition (Figure 6A), the IFR grouped condition (Figure 6C), the ISR ungrouped condition (Figure 6E) and the ISR grouped condition (Figure 6G); and the right-hand panels are the model simulations from the IFR ungrouped condition (Figure 6B), the IFR grouped

condition (Figure 6D), the ISR ungrouped condition (Figure 6F) and the ISR grouped condition (Figure 6H).

Figure 7. The proportion of CRP Lag+1 transitions within and across group boundaries. The left-hand panels are the data from the IFR ungrouped condition (Figure 7A), the IFR grouped condition (Figure 7C), the ISR ungrouped condition (Figure 7E) and the ISR grouped condition (Figure 7G); and the right-hand panels are the model simulations from the IFR ungrouped condition (Figure 7B), the IFR grouped condition (Figure 7D), the ISR ungrouped condition (Figure 7F) and the ISR grouped condition (Figure 7H).

Figure 1

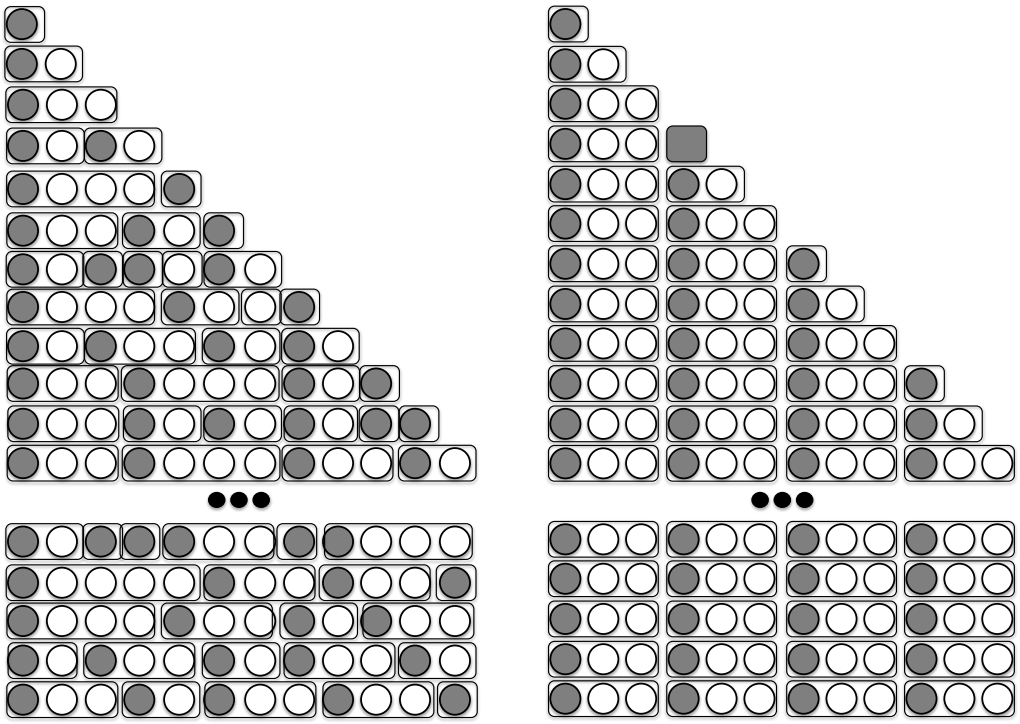


Figure 2.

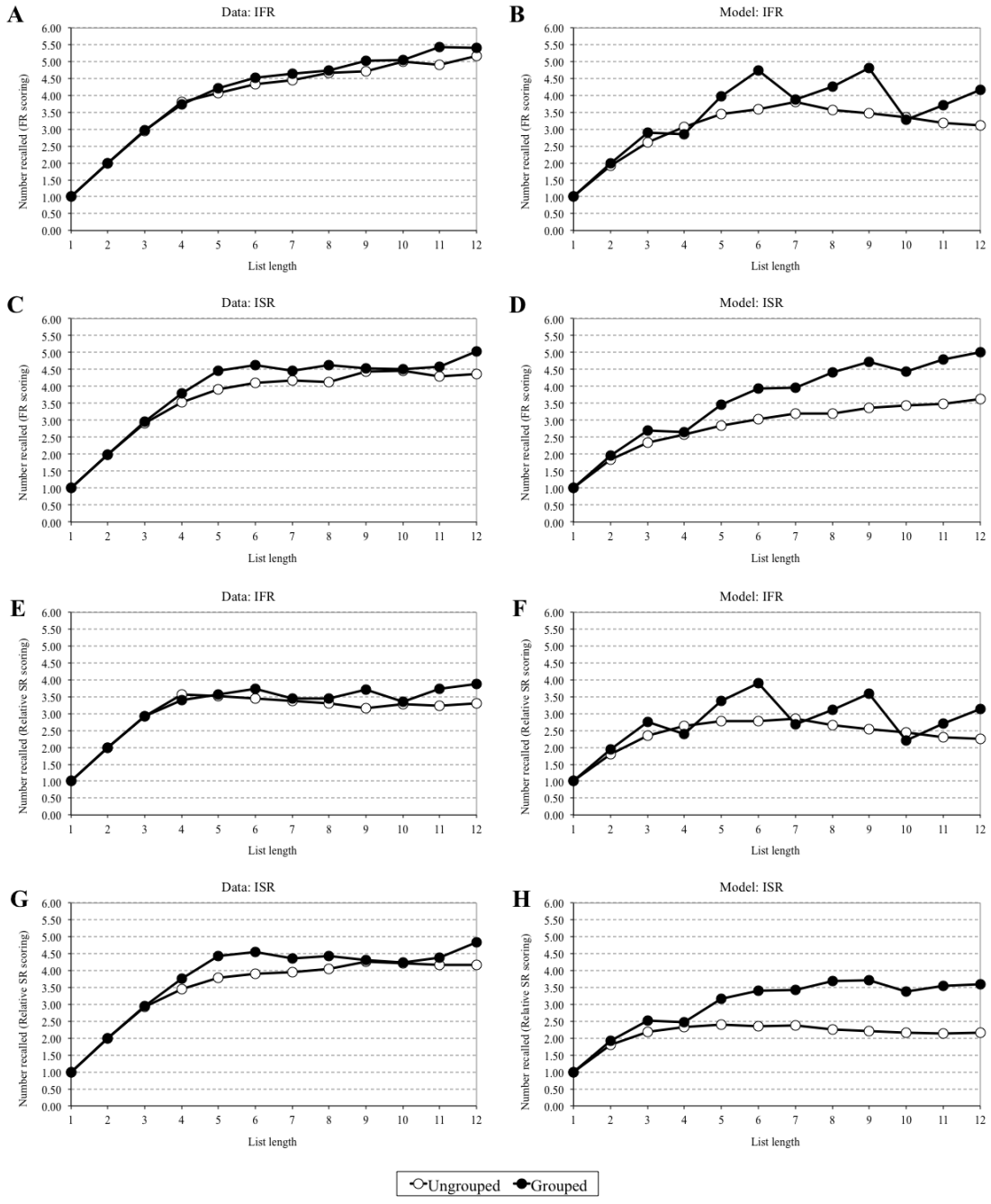

A
Data: IFR
B
Model: IFR
C
Data: ISR
D
Model: ISR
E
Data: IFR
F
Model: IFR
G
Data: ISR
H
Model: ISR
Number recalled (FR scoring)
Number recalled (Relative SR scoring)
List length
Ungrouped
Grouped

Figure 3.

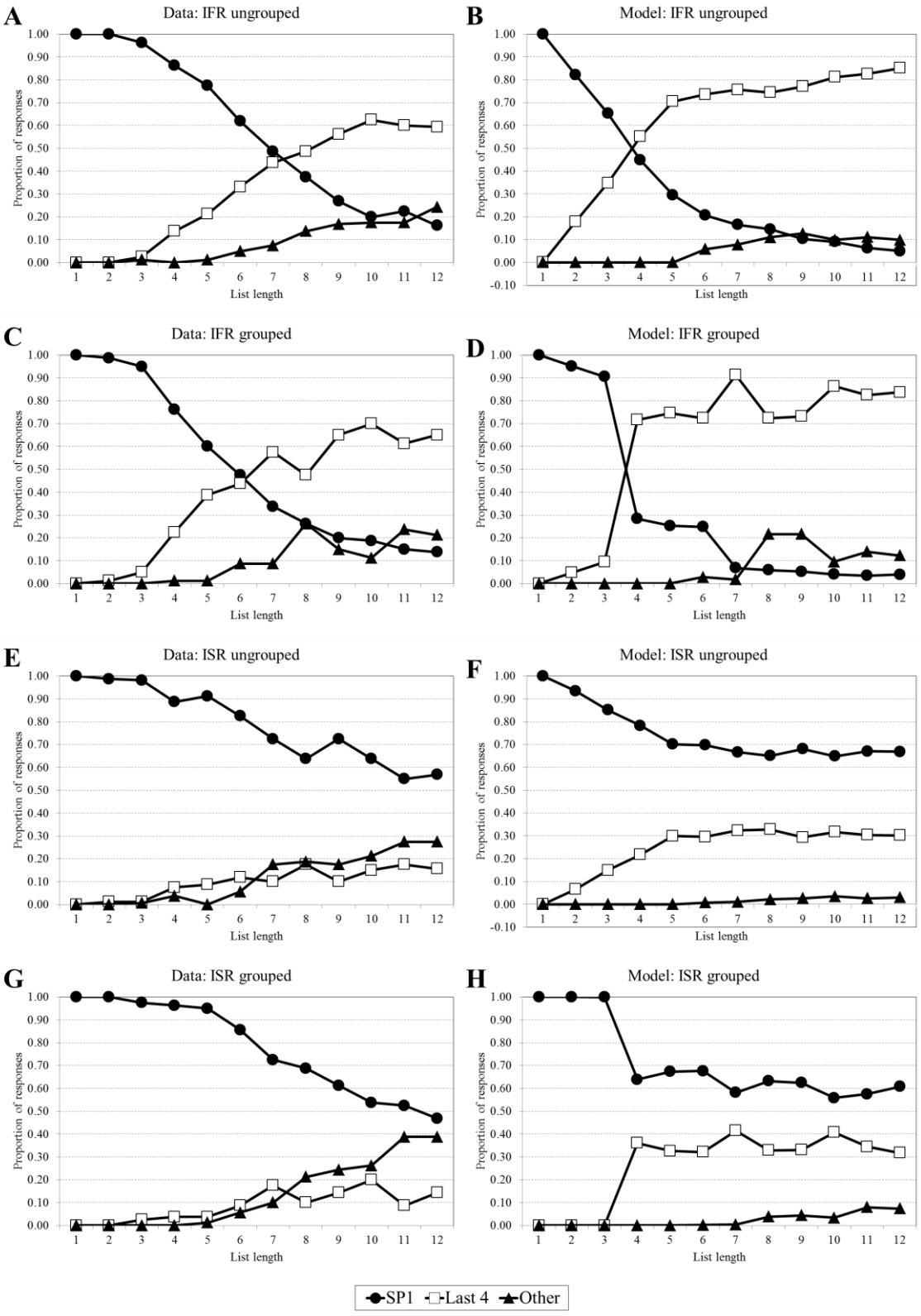

A
Data: IFR ungrouped
B
Model: IFR ungrouped
C
Data: IFR grouped
D
Model: IFR grouped
E
Data: ISR ungrouped
F
Model: ISR ungrouped
G
Data: ISR grouped
H
Model: ISR grouped
Proportion of responses
List length
SP1
Last 4
Other

Figure 4.

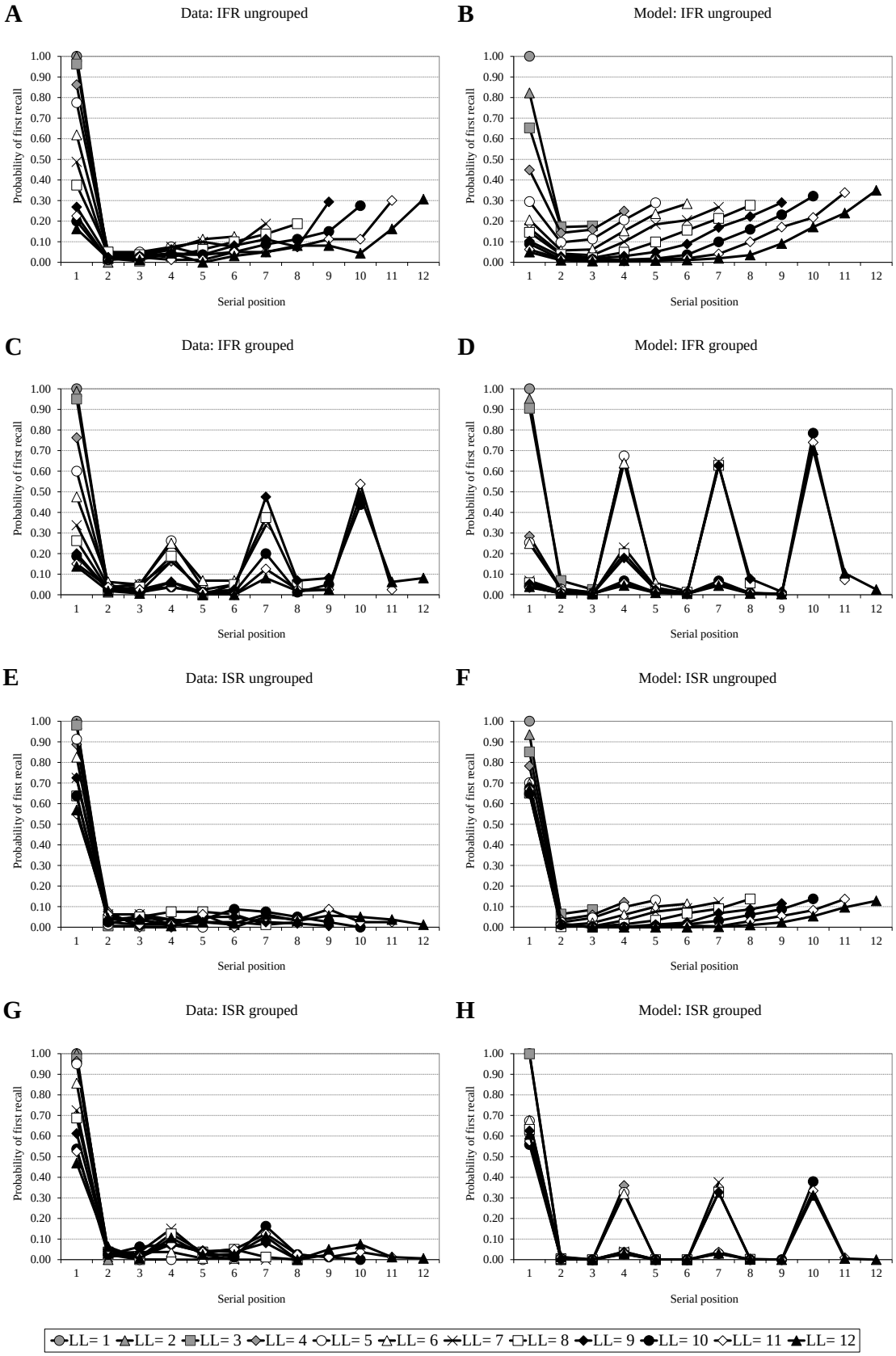

A
Data: IFR ungrouped
B
Model: IFR ungrouped
C
Data: IFR grouped
D
Model: IFR grouped
E
Data: ISR ungrouped
F
Model: ISR ungrouped
G
Data: ISR grouped
H
Model: ISR grouped
Probability of first recall
Serial position
LL= 1
LL= 2
LL= 3
LL= 4
LL= 5
LL= 6
LL= 7
LL= 8
LL= 9
LL= 10
LL= 11
LL= 12

Figure 5.

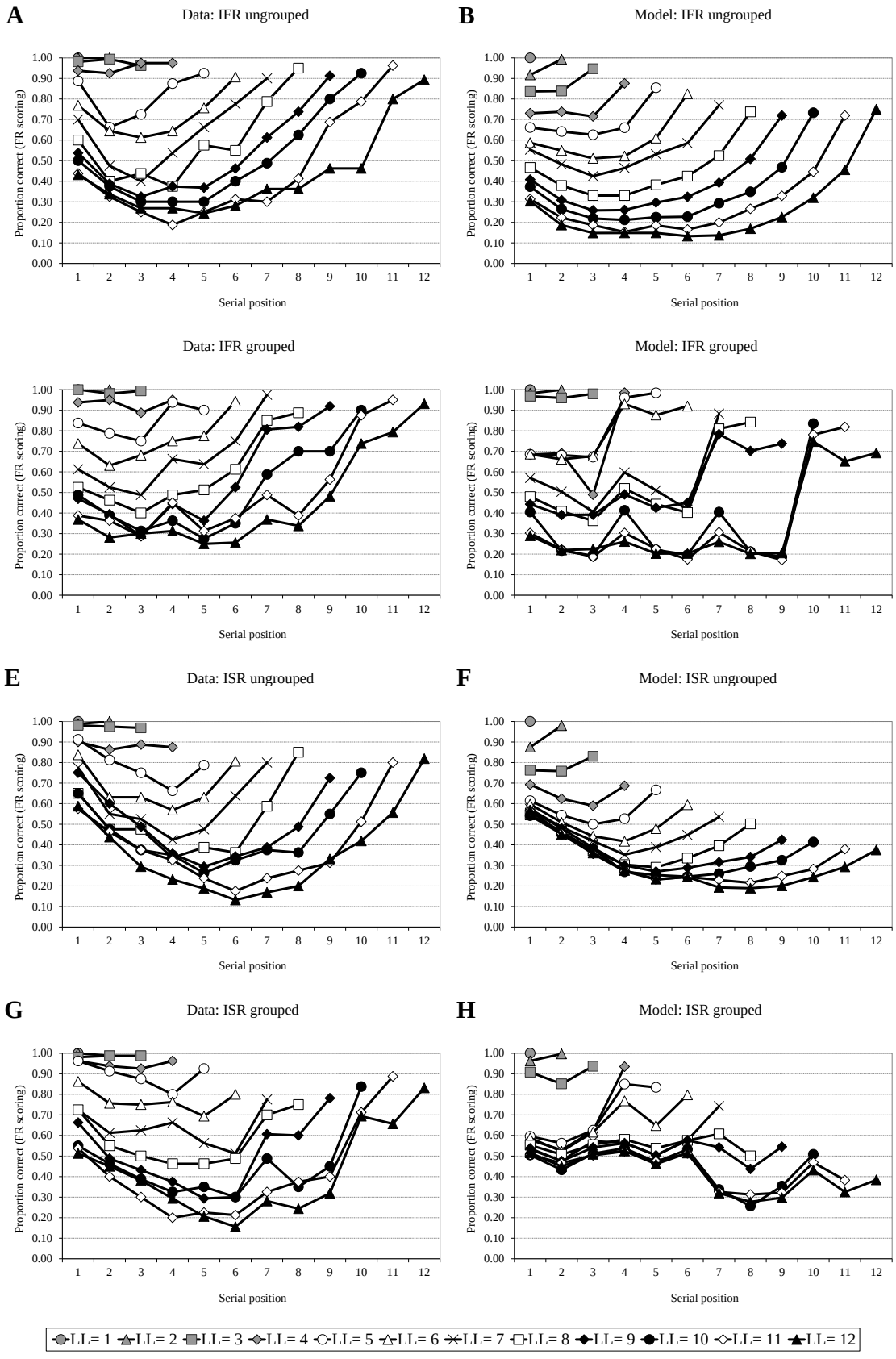

A
Data: IFR ungrouped
B
Model: IFR ungrouped
Data: IFR grouped
Model: IFR grouped
E
Data: ISR ungrouped
F
Model: ISR ungrouped
G
Data: ISR grouped
H
Model: ISR grouped
Proportion correct (FR scoring)
Serial position
LL= 1
LL= 2
LL= 3
LL= 4
LL= 5
LL= 6
LL= 7
LL= 8
LL= 9
LL= 10
LL= 11
LL= 12

Figure 6.

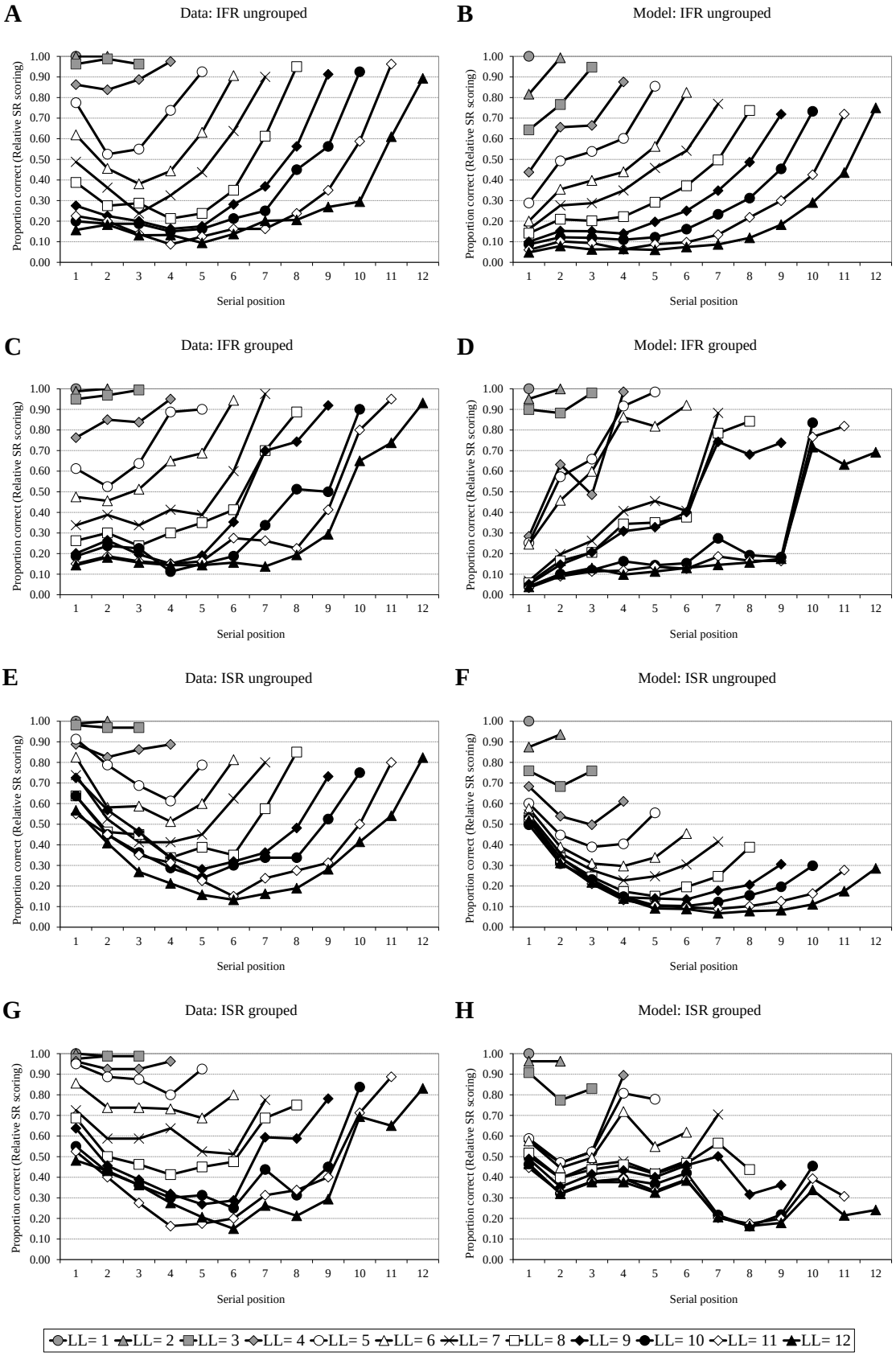
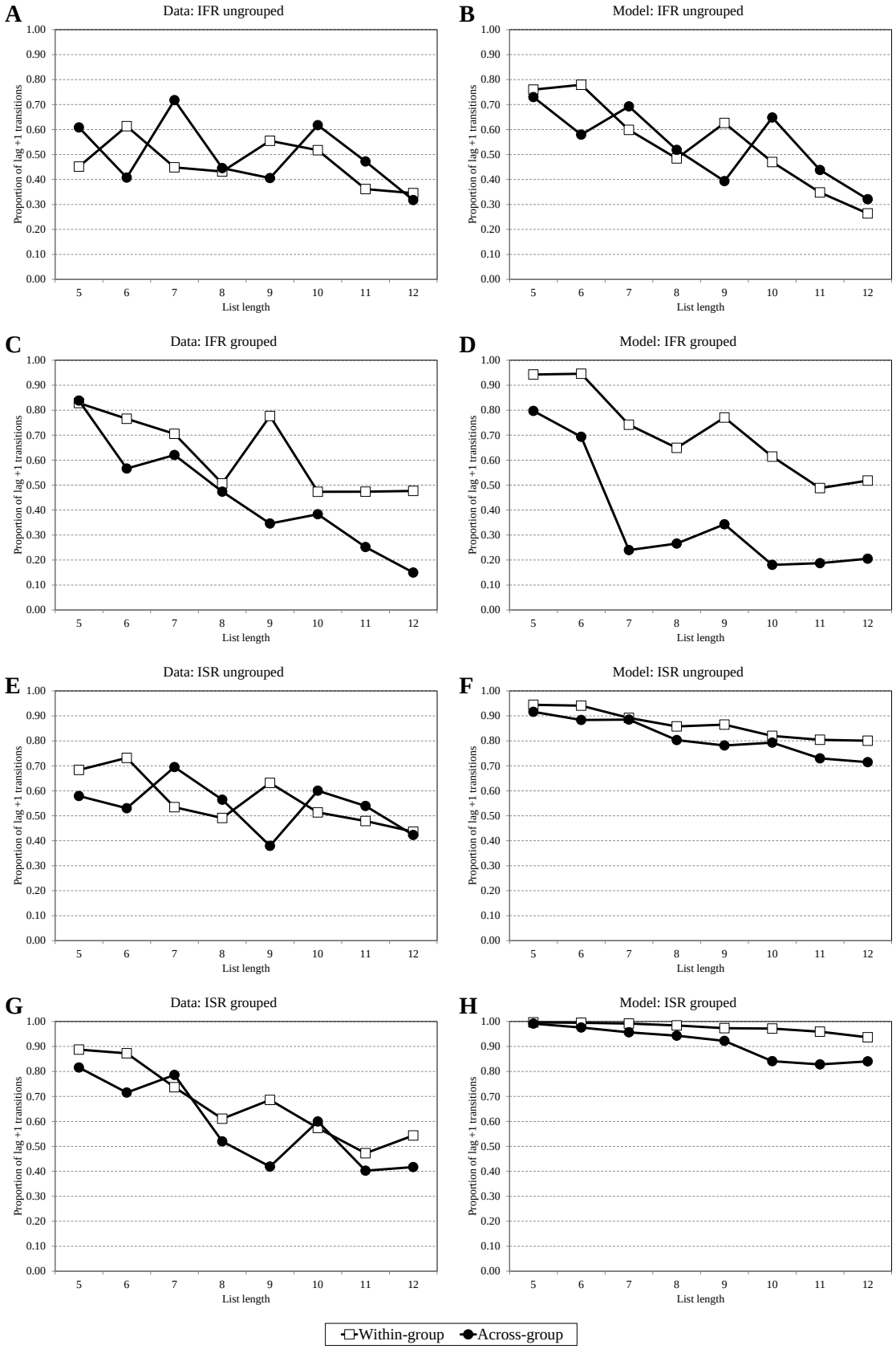

A
Data: IFR ungrouped
B
Model: IFR ungrouped
C
Data: IFR grouped
D
Model: IFR grouped
E
Data: ISR ungrouped
F
Model: ISR ungrouped
G
Data: ISR grouped
H
Model: ISR grouped
Proportion correct (Relative SR scoring)
Serial position
LL= 1
LL= 2
LL= 3
LL= 4
LL= 5
LL= 6
LL= 7
LL= 8
LL= 9
LL= 10
LL= 11
LL= 12

Figure 7.


A
Data: IFR ungrouped
B
Model: IFR ungrouped
C
Data: IFR grouped
D
Model: IFR grouped
E
Data: ISR ungrouped
F
Model: ISR ungrouped
G
Data: ISR grouped
H
Model: ISR grouped
Proportion of lag +1 transitions
List length
Within-group
Across-group