Human order memory: insights from the relative-order task

by

Yang Liu

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Psychology
University of Alberta

Abstract

In our daily activities, whether it is to remember a phone number, a recipe or a movie plot, remembering order information is crucial. The most common way to study order memory is serial recall, where participants are asked to recall a study list in the order that the list was presented. An alternative approach is the relative order task, where participants are given two items from a study list and asked to judge which one came earlier or later. In judgements of temporal order in short lists, a congruity effect is found: asking "which item came earlier" versus "which item came later" reverses search direction. The finding of a congruity effect in short lists led to a series of questions of whether the same congruity effect could be generalized to list of different types and whether behaviour data from relative order judgements could be accounted for by memory theories developed to explain serial-recall data. In this dissertation I report results from a series of studies focusing on the congruity effect beyond short lists and the response time measure, and relating theories of serial recall to theories of comparative judgements. Specifically, those studies report that the congruity effect generalizes to longer lists, the English Alphabet, and grouped lists, as well as the error rate measure. The generality of the congruity effect suggests current versions of order memory models need further assumptions to account for this effect. In addition,we report that grouping effects on relative order judgements are compatible with a positional coding model with two-level hierarchies. The comparison to the effects of grouping on serial recall suggests how relative order judgements and serial recall may share the same cognitive mechanisms. Together, these behavioural results further establish the generality of the congruity effect, bridge order memory theories based on relative order judgement and serial recall data, and set new constraints on future memory model development.

Preface

This thesis is an original work by Yang S. Liu. All research projects contributing to this work received ethics approval from the University of Alberta Research Ethics Board. Project Name "Organisation and Retrieval Timecourse of Human Memory", No. Pro00009760.

Chapter 2 of this thesis has been published as Liu, Y. S., Chan, M., & Caplan, J. B. (2014), "Generality of a congruity effect in judgements of relative order". *Memory & Cognition, 42*(7), 1086-1105. I was involved for concept formation, experiment paradigm implementation, data collection, data analysis and interpretation, and the manuscript composition. M. Chan was involved with concept formation and data collection. J. B. Caplan was involved with concept formation and manuscript composition.

Chapters 3, 4 and 5 of the thesis has not been published elsewhere. I was responsible for concept formation, experiment paradigm implementation, data collection, data analysis and interpretation, and the manuscript composition. A. Gupta assisted in data collection of Chapter 5. J. B. Caplan was involved with concept formation and manuscript composition of Chapter 3, 4 and 5.

Acknowledgements

The work of my dissertation was not possible without the tremendous support of people at the University of Alberta, my family, and the organizations funding the research. First and foremost, I would like to thank my supervisor Dr. Jeremy Caplan for his excellent mentorship, helpful advice, enormous support and encouragement, the works in this dissertation was not possible without Jeremy's guidance. I would also like to thank my PhD. supervising committee Dr. Norman Brown and Dr. Weimin Mou for good advice on my work. I would like to express my appreciation to Michelle Chan, and Aditi Gupta for their help with running experiments. Particularly I want to thank my fellow lab members Christopher Madan, Yvonne Chen, Leanna Cruikshank and Kenichi Kato for their friendship, mindful discussions and generous support. I would also like to thank my parents and my wife Qian Cheng for their emotional support throughout my PhD program. Finally, I would like to appreciate the funding support made my research possible, from the Alberta Ingenuity Fund, the Natural Sciences and Engineering Research Council of Canada (NSERC), the Psychology Department of the University of Alberta and Queen Elizabeth II Scholarship from the University of Alberta.

Sincerely,

Yang S. Liu

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Δ BIC	Change in the Bayesian Information Criterion
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
ISI	Inter-Stimulus Interval
JOR	Judgements of Relative Order
LME	Linear Mixed Effects
OSCAR	OSCillator-based Associative Recall
SAT	Speed Accuracy Tradeoff
SFGT	Slow-Fast Guessing Theory
SIMPLE	Scale Independent Memory Perception and Learning
TODAM	Theory of Distributed Associative Memory

Chapter 1

Introduction

Order memory is a key aspect of our daily activities. You may remember the individual digits of a phone number, the ingredients of a recipe, but it is the order of digits and steps of cooking that makes the memory meaningful. The works in this dissertation are focused on better understanding of how human memory works and producing empirical data that can potentially challenge current models and constrain, as well as guide, future development of mathematical models of memory. This dissertation focuses on the behavioural paradigm of judgements of relative order (JOR), where participants are asked to judge relative order between two items from a study list. To understand where this research stands relative to prior research, we next review: 1) Methods and theories of order memory 2) Grouping effects and hierarchical positional coding 3) Congruity effects on comparative judgements 4) JOR as comparative judgement.

1.1 Methods and theories of order memory

The most common method of studying temporal order memory is serial recall, that is, asking people to recall a list they just learned, in its original order. For instance, if you just learned a telephone number "7091238", serial recall would require an exact output of "7091238". If you, instead, recalled "7901238", you made an order error switching the positions of "0" and "9". A major limitation of serial recall is that order memory depends on remembering the individual items from the list. A long list length further exacerbates this problem as the chance of remembering individual items drops rapidly with increasing list lengths (Murdock, 1974). Despite its limitations, serial recall has a long history of empirical research (e.g., Murdock, 1974; Neath & Surprenant, 2003) and serial recall data

have dominated the development of order memory theories. The serial-recall procedure has provided a large number of empirical findings that have been useful in selecting and constraining models of order-memory. We will present a complementary procedure that further contribute to prior works based on serial recall.

An alternative method to study order memory is asking for the relative order of two list items. Taking the telephone number example, a relative order judgement task could ask "Which digit is more recent? '0' or '9' "? Researchers asking which item is more recent have referred to this task as "recency judgement" (Hacker, 1980; Hockley, 1984; Muter, 1979; Yntema & Trask, 1963). Asking relative order of two digits in a telephone number is rather an odd task to do in our daily life, as all other digits need to be remembered correctly to successfully telephone someone. However, it is perfectly appropriate when the absolute positions of individual items do not matter. For instance, "who left the party earlier? Weimin or Norman?" It could be that twenty people left before Norman, but with this phrasing of the question, that is irrelevant; all we care about is whether Weimin left earlier or later. Thus, both serial recall and JOR studies should contribute to our understanding of human order memory. The wording of the question is crucial for the JOR procedure, as we will discuss later. Until recently, only a handful of studies looked at temporal order memory using the JOR procedure using a wording like "which item is more recent?" (Hacker, 1980; Hockley, 1984; Muter, 1979; Yntema & Trask, 1963) and empirical results from recency judgements have not been considered as benchmark findings that models need to explain. The results from recency judgements (Hacker, 1980; Muter, 1979) suggested the more recent item from the pair dominates the response time, and the behavioural patterns suggested participants use a backward self-terminating search strategy to look for the most recent item. To perform a backward-self terminating search, participants start from the end of the list and scan towards the beginning of the list. As soon as the most recent item is found, the search process stops. Backward self-terminating search therefore predicts response time or error rate decreases as the later probe position decreases. In Figure 1.1, we show a schematic of backward self-terminating search, showing the position of earlier item position does not influence response time, yet response time increases as later item position decreases. To our knowledge, only a handful of memory models have been directly applied to recency-judgements data. Hacker (1980) modelled his recency-judgements data assuming the backward self-terminating search mechanism. The first recalled item matched with one

of the two probes is assumed to be the more recent item. Error responses are made when the recalled item matches the wrong probe, or when no match is found and participants guess at 50% accuracy. Hacker's (1980) self-terminating search model successfully fit behavioural data of recency judgements. McElree (1996) also fitted their data to Hacker's model, and found self-terminating search could adequately account the data pattern. Following this success of a backward self-terminating search model, the OSCillator-based Associative Recall (OSCAR) model (Brown, Preece, & Hulme, 2000) also successfully modelled Hacker's (1980) data. In this model, items are assumed to be associated with the state of an internal context signal (activation values of a bank of sine-wave oscillators), and retrieval of items requires re-instatement of the context. OSCAR models judgements of recency results by first using the end of the list context to probe for retrieving the strongest associated item, then the strongest associated item will be compared with the two probes. When there is a match to a probe item, the probe item is chosen as the more recent. If no match is found, the next strongest item is used for this comparison, and this process repeats until a match is found. This process is essentially implementing a backward self-terminating search. However, as an alternative mechanism to sequential self-terminating search, OSCAR is also compatible with a direct item access by probing with the corresponding context, which we broadly classify as a direct access model.

For decades, no one thought that asking a different question "which item came earlier?" would make a difference, and no explicit comparison between instructions were made. This question is asking for the same information as "which item came more recently?" or put into other words "which item came later." There could be subtle differences between the "later" and "recent" questions, as the "recent" instruction cues the participants to set the point of reference to the current time, where the "later" instruction does not suggest a point of reference. Although no direct comparisons have been conducted between the "later" and "recent" instruction, both instructions show the same qualitative behavioural patterns as we will describe in Chapter 2 (Liu, Chan, & Caplan, 2014). Chan, Ross, Earle, and Caplan (2009) tested both "earlier" and "later" instructions on lists of consonants (list lengths 3, 4, 5, and 6), and confirmed how we ask the question matters. When asked "which item came earlier", the results are consistent with a forward self-terminating search response time pattern, rather than the classical backward self-terminating search pattern found in the "recency" or "later" instruction. A direct consequence of this finding is that order-

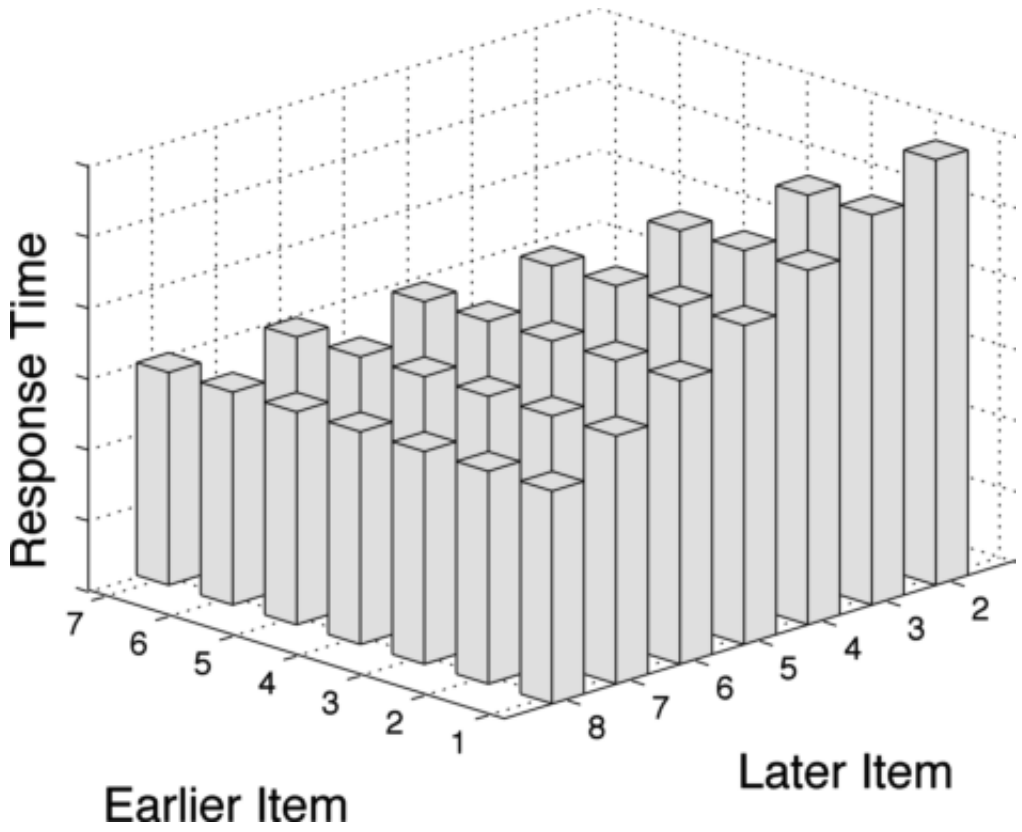


Figure 1.1: Schematic depictions of hypothesized backward self-terminating search. Response time is plotted as a function of both the earlier probe-item's serial position ("Earlier Item") and later probe-item's serial position ("Later Item").

memory models can no longer assume a backward search strategy to account for all recency judgements results. Relative order judgements are not limited to the "recency" instruction, thus we broadly refer to this paradigm as JOR, which we use as an acronym for "Judgments of relative ORder".

The works in this dissertation build on Chan et al.'s (2009) findings, as there is very little we know about how instruction affects JORs across different conditions. We call this instruction effect a congruity effect, a more general term to describe the crossover interaction between instructions and the list serial positions, where the "earlier" instruction selectively facilitates judgements for probes toward the beginning of the list and the "later" instruction selectively facilitates judgements for probes toward end of the list. In Chapter 2 (Liu et al., 2014), we ask whether instruction has the same effect on longer lists. In this dissertation, we use the term "subspan" and "supraspan" to refer to short lists where response accuracy is at ceiling, and long lists where response accuracy is below ceiling, respectively. Our goal is to investigate whether the congruity effect could be found when the error rate is high, without assuming there are different time scales. We further asks whether this effect affects access speed only or also affects accuracy. In two separate experiments, we test both word lists (list lengths 4, 6, 8, and 10) and consonants lists (list lengths 4 and 8) with both the "earlier" and "later" instruction, and find instruction has similar effects on the supraspan lists after controlling other factors typically found on supraspan recency judgements. In addition, instructions have similar effects on error rates as on response times. The results suggest that for supraspan lists, the "earlier" instruction does not induce a forward self-terminating search behavioural pattern; however, the "earlier" instruction facilitates response time and reduces errors towards the beginning of the list, and the "later" instruction facilitates response time and reduces errors towards the end of the list. Thus, we have confirmed the congruity effect on supraspan lists. In Chapter 3, we ask whether the congruity effect could have been found on a well learned long list, and choose the English alphabet as the target list. We find the response time congruity effect on alphabetical order judgements is similar to the supraspan JOR results, and further find an error rate congruity effect after controlling speed-accuracy tradeoffs.

We wonder whether the congruity effect could be explained by the recall direction in serial recall. Data from Chapter 2 (Liu et al., 2014) suggest scanning direction differences can not explain the congruity effect, because the JOR behaviour in supraspan lists is dominated by

an overall recency effect. The "earlier" instruction may induce participants to process the list more like forward serial recall, whereas the "later" instruction may induce participants to process the list more like backward serial recall. In Chapter 4 we test this hypothesis by giving participants a 9-item consonant list and asking them to either recall the list in forward or backward order. We find that when accuracy is plotted as a function of the recall output order, those output position plots for both the forward and the backward serial recall are qualitatively similar. In addition, we also manipulate the presentation of the list. We test whether grouping the list into groups of 3s by introducing a temporal gap versus a ungrouped list with even inter-item interval could affect serial recall. This experiment is designed to be interpreted in conjunction with results from Chapter 5, where we use the same presentation methods and materials from Chapter 4, and test with JOR instead of forward and backward serial recall. The matched presentation methods and materials allow us to compare the serial position effects between the JORs and serial recall. In addition, we also test whether the congruity effect generalizes to the grouped list. We find the JOR congruity effect cannot be explained by participants recalling the list in opposite directions, and the congruity effect generalizes to grouped lists. We will further discuss the grouping research in the next section.

1.2 Grouping effects and hierarchical positional coding

One of the most common ways of plotting serial recall behaviour is the serial position curve (SPC), which plots accuracy as a function of study positions. Serial position curves show non-smooth accuracy transitions between groups, suggesting participants spontaneously organize information in groups (e.g., Martin & Noreen, 1974; Madigan, 1980; Jou, 2011; Wickelgren, 1967; Ryan, 1969a). Grouping could also be induced experimentally. The effects of grouping have been extensively studied (Brannon, 1997) and drive memory theories based on serial recall results (e.g., Henson, 1998).

Grouping enhances overall recall accuracy, and, more specifically, induces a so-called "scalloped" pattern in serial position curves (Henson, 1998; Lee & Estes, 1981; Maybery, Parmentier, & Jones, 2002; Ng & Maybery, 2002, 2005; Wickelgren, 1967; Ryan, 1969a). The "scalloped" pattern is characterized by a mini serial position curve for each group, each shows within-group primacy and recency effects. In addition, grouping affects the

kinds of errors participants make in very specific ways (Ryan, 1969a, 1969b). In ungrouped lists, the most common order error is adjacent transposition (Lee & Estes, 1977), that is when an item is remembered in its adjacent serial positions. Compared to ungrouped lists, adjacent transposition errors for grouped lists are lower overall, and especially lower when the transposition is across two groups. In addition, grouping introduces one type of transposition error, that is when an item is recalled in the wrong group but with the correct within-group position (Henson, Norris, Page, & Baddeley, 1996; Farrell & Lewandowsky, 2004). Those errors were termed interposition errors (Henson et al., 1996) .

Latency data also carries information about how participants remember grouped lists. Participants were slower to recall the first item within each group, followed by much faster response times for the remaining within-group items (Maybery et al., 2002; Ng & Maybery, 2002, 2005; Farrell & Lelièvre, 2012; Thomas, Milner, & Haberlandt, 2003). Thus, recall is faster within-group than between-group. The recall-latency data provide complementary evidence from accuracy results. More specifically, grouping increases recall speed for the first item in each group, but reduces recall speed for other items from the list.

The results from forward recall are consistent with theories assuming memory is represented by a hierarchical or multidimensional structure. This class of positional coding theories generally considered the list items are coded by its position and recalling a item requires probing memory directly with the position code, and groups can also be represented by a position code relevant to the whole list, or within individual groups (J. R. Anderson & Matessa, 1997; Hurlstone, Hitch, & Baddeley, 2014). The positional coding theory is further reviewed in Chapter 4.

Surprisingly, how backward recall fits in with the positional coding theories is poorly understood. Thomas et al. (2003) and Haberlandt, Lawrence, Krohn, Bower, and Thomas (2005) have suggested backward recall is achieved by a scan-and-drop strategy (Conrad, 1965), where multiple forward recalls are performed starting from the beginning of the list, and when the target item is recalled, it is dropped from the search set. For example, for a list, A, B, C, D and the backward recall task, the scan-and-drop strategy could be consist of a forward scan of ABCD and output D. After D is recalled from the list, it is dropped from the set and recalling C only requires a forward scan of ABC. This process continues until the full list is recalled. We wonder whether grouping in backward serial recall would provide the same support for positional coding models. Farrell and Lelièvre (2012) made the

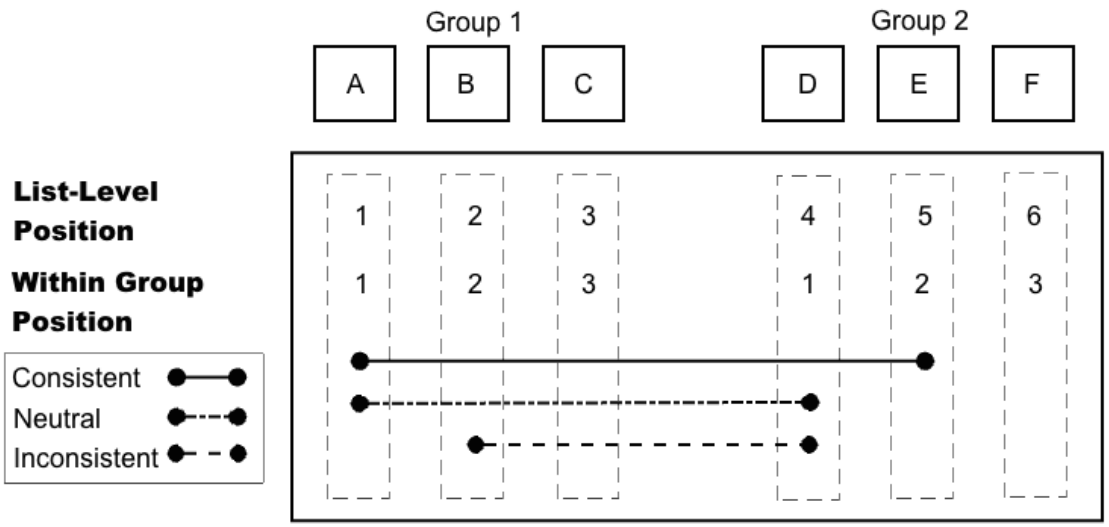


Figure 1.2: Schematic of two-level position coding for list ABCDEF, with ABC in the first group and DEF in the second group. The black dots connected by lines are examples pairs when relative order judgements based on the two-level position codes are Consistent, Neutral, or Inconsistent.

assumption that group-level position code is coded in forward order. We therefore also ask whether within-group position is directly retrievable in backward order. In Chapter 4 we study whether empirical data of backward serial recall could show evidence of the scan-and-drop strategy, and ask how can backward recall in a grouped list constrain assumptions made by positional coding models. We find no evidence supporting the scan-and-drop strategy and our results suggest group-level position codes could be retrieved in backward order.

In addition, we wonder whether the JOR paradigm could provide converging evidence to support hierarchical positional coding theories. The positional coding theories could explain order judgements by determining the relative order of position codes. When a list is grouped into groups, there are position codes for list items, as well as position codes at each individual groups. For example, for the consonant list JBLCDFK, the list-level position code is 1234567, with each position corresponding with each consonant sequentially. More specifically, "J" has a position code of 1, "B" has a position code of 2, "L" has a position code of 3 and so on. "K" is the last item in the list and has a position code of 7. When we ask "which item came earlier/later? 'B' or 'F' ", one could compare the position codes of "B" and "F" and make a correct judgement if retrieved position codes are in the correct order. If order memory is organized by hierarchical position codes, one can expect position codes at each individual

group. For example, if the consonant list is organized in two groups "JBL" and "CDFK", the group-level position code would be "123" and "1234" for each group respectively. More specifically, in the first group "JBL", "J" has a group-level position code of 1, "B" has a group-level position code of 2, "L" has a group-level position code of 3, in the second group "CDFK", "C" has a group-level position code of 1, "D","F" and "K" have group-level position code of 2, 3, and 4 respectively. Sometimes judgements based on the group-level position codes predict the same results as judgements based on the list-level position codes, and we call this the consistent case. For example, "B" and "F" have group-level position code 2 and 3 respectively, one can infer "B" is earlier than "F". However, sometimes group-level position codes predict the opposite answer from list-level position codes, and we call this the inconsistent case (e.g., "B" and "C"). Figure 1.2 is a schematic of the two-level position codes with examples of Consistent, Neutral and Inconsistent pairs. If a hierarchical organization of position codes exists, the inconsistent case should impair accuracy. In Chapter 5, we test both grouped and ungrouped lists using the JOR procedure, and find that the consistent case had better accuracy than the inconsistent case. In addition, we find grouping enhances the overall accuracy, at a cost of slower response times. The results generally support the predictions of hierarchical positional coding models.

1.3 Congruity effects on comparative judgements

As early as the 1900s, Cattell (1902) asked participants to discriminate between properties of two stimuli, such as which card has a higher luminance, and found a systematical relationship between perception and stimuli intensity. This two-alternative forced choice task, where properties of the stimuli are judged along a continuum, is referred as comparative judgements. Since its early 20th century introduction, this line of research has been expanded to perceptual domains, including weights of objects (Masin, 1995; Paivio, 1975), loudness (Holyoak & Patterson, 1981), pitch (Audley & Wallis, 1964; Banks & Root, 1979), length and horizontal extent (Petrusic, 1992; Petrusic & Baranski, 1989). Following the same logic, this two-alternative comparison task extended to the symbolic domain that requires memory, such as judging differences between size properties of a semantic term (Banks, White, Sturgill, & Mermelstein, 1983; Banks & Flora, 1977; Cech, 1995; Cech & Shoben, 2001; Shoben, Cech, Schwanenflugel, & Sailor, 1989; Shoben & Wilson, 1998), numbers

(Duncan & McFarland, 1980), alphabet (Jou, 1997; Jou & Aldridge, 1999), geographical locations (Maki, 1981), ranked order (height; e.g., Jou, 2011), temporal order (Chan et al., 2009; Hacker, 1980; Marshuetz, 2005; McElree & Dosher, 1993; Muter, 1979; Liu et al., 2014; Yntema & Trask, 1963).

From the perspective of comparative judgements, one can conceptualize order memory as judgements of rankings along a linear continuum (Jou, 2011), and for order memory, the relevant dimension is time (Brown, Neath, & Chater, 2007). The JOR procedure could be broadly classified under comparative judgements, if we consider temporal order judgement as a special case of magnitude judgement (Jou, 2011) along the dimension of time (Brown et al., 2007), and mechanisms underlying comparative judgements might be applicable to JORs.

Across the wide range of stimulus types and dimensions, three very robust phenomena have been found for comparative judgements: a) an end effect, also termed as a serial position effect. The end effect is characterized by an inverted U-shaped response time or error rate curve, with better performance for items at either extreme end of the list, and worse performance at the middle of the list; b) a distance effect, characterized by faster response time and lower error rate when the actual or symbolic distance between the probe items increases (e.g., when asked which animal is larger between rhino and rabbit, the response time is faster than the same judgement between rhino and elephant); c) a congruity effect, characterized by a faster response time and lower error rate when the wording of the question is congruent with the probe on a relevant dimension (e.g., when asked which animal is larger between rhino and elephant, the response time is faster than the same judgement between cat and rabbit). The "larger" question is congruent with the larger size of the rhino and elephant, thus enhancing performance.

As we reviewed in the first section, the comparative judgement paradigm has been largely overlooked by serial order researchers and modellers (but see Brown et al., 2007; Jou, 2011). The studies using the judgements of recency task, where participants were asked "which item is more recent?" between two items, have showed evidence of a U-shaped serial position effect, and a distance effect (Hacker, 1980; Hockley, 1984; Muter, 1979; Yntema & Trask, 1963), but because only one question is asked, the congruity effect is not tested. It is possible that the same congruity effect could be found in JORs, and models explaining the congruity effect for comparative judgements should be considered.

Some researchers have proposed the congruity effect could be explained by its semantic codes (Banks, 1977). According to the semantic coding model (Banks, 1977), the congruity effect is caused by a semantic code's match or mismatch with the instruction. For instance, when elephant and rhino are coded as "big+" and "big"', the instruction "choose larger" matches with the semantic code. Thus, no further translation of semantic code is required. If, however, a "choose smaller" instruction is given, the original code needs to be translated to a matching code of "small" and "small+" ("+" means more). A related mechanism to semantic coding theory is the semantic interference theory (Banks & Root, 1979), which assumes that the semantic attribute of the probe can interfere with the semantic meaning of the question (e.g., "shortness" of a short item may interfere with "taller" instruction's tallness).

A few explanations have been proposed to account for the congruity effect without considering semantic codes. One of these is the the reference point theory (Holyoak, 1978; Jamieson & Petrusic, 1975). The reference point refers to points on a continuum that other values can take as a reference, and usually depends on the task-instruction (e.g., the question "Which number is larger than 5?" implies the reference point is 5, and all other numbers would be compared to 5). The extreme members of a continuum are considered the default reference points when no specific reference point is provided. The model assumes each judgement is made by comparing both stimuli to a reference point and the ratio of differences between stimuli and reference point determines the congruity effect. Another explanation is the end-inwards search model (Woocher, Glass, & Holyoak, 1978; Holyoak & Patterson, 1981), where participants are assumed to scan from either end of the list towards the center of the list and looking for a match with the probe items. The congruity effect is explained by the assumption that search from the congruent end of the list is relatively faster. However, the end-inwards search models may not be sufficient to explain comparative judgements. The end-inwards scanning model predicts when both items are near the center of the list, it takes longer to find the probe items regardless which direction the self-terminating search started. This mechanism predicts that if the task can be made sufficiently easier, the search speed should be faster, and the distance effect should be reduced or eliminated. Holyoak and Patterson (1981) found that lowering task difficulty does not change the distance effect for probes near the middle of the list.

Instead of directly altering the search direction, the congruity effect can also be explained

assuming differential bias caused by the specific instruction. For example, a "large" bias produce faster processing of large items, and a "small" bias produce faster processing of small items. Models based on differential bias have been applied to empirical data (e.g., Birnbaum & Jou, 1990). We will further discuss how concepts from comparative judgements could help us understand order memory in Chapter 6.

1.4 JOR as comparative judgement

As we discussed in the last section, it is important to consider whether JORs could be understood as a special case of comparative judgements. In Chapter 2 (Liu et al., 2014), Chapter 3 and Chapter 5, we demonstrate broad boundary conditions for the JOR congruity effects, as well as replicate the primacy/recency effect and the distance effect. The shared benchmark effects and two-alternative forced choice testing methods suggest the JOR tasks could be understood as a comparative judgement task along the continuum of time.

However, it is important to also consider the differences between the JOR and traditional comparative judgemental tasks. The comparative judgements literature typically measures response times, where the list is usually trained to a high accuracy criterion. The ceiling accuracy or trained accuracy criterion restricts the interpretation of the response time data, as the response times and error rates may trade off. In Chapter 2 (Liu et al., 2014) and Chapter 5, participants learn the list sequentially with no list relearning, and both accuracy and response times are measured. Unlike comparative judgements, researchers who use serial recall as a paradigm focus on error rate measures primarily, and memory theories developed for serial recall typically addresses empirical data from error rates. Therefore, the error rate results from the JOR paradigm are essential to connect the JOR and serial recall paradigm.

1.4.1 Congruity effects

The JOR paradigm already showed a U-shaped serial position effect and a distance effect, when participants were asked to choose the more recent item from the probe (Hacker, 1980; Hockley, 1984; Muter, 1979; Yntema & Trask, 1963). Thus, the finding of a congruity effect would suggest the JOR paradigm could be understood as a comparative judgement task. In Chapter 2 (Liu et al., 2014) Chapter 3 and Chapter 5, we show the congruity effect and other benchmark effects are consistently replicated across broad boundary conditions.

A novel result from the JOR paradigm is the error rate congruity effect. An error rate congruity effect has rarely been found in comparative judgements, with a few exceptions (Petrusic, 1992). The error rate congruity effect suggests the underlying process of relative order judgement is affected by both the availability and the quality of information. The demonstration of the error rate congruity effect is also essential for challenging memory models that focuses on modelling error rate results.

When we consider the JOR task as a comparative judgement task, It is worth noting that the congruity effects observed are slightly different from the semantic congruity effects. For the semantic congruity effect, participants are asked to judge between two items that were both semantically congruent or incongruent with the the instruction. For example, elephant and rhino are both large animals, and this pair is congruent with the "choose larger" instruction, and incongruent with the "choose smaller" instruction. For the JOR paradigm, we have tested all possible probe combinations, where it is possible to have a probe with one item from the earlier part of the list, and another item from the later part of the list. We suggest the underlying congruity effect is the same between the semantic lists and temporal ordered lists, and the mechanism is should not be limited by the semantic coding theory. Testing all possible combinations of probe pairs allows us to interpret the results with the distance effect, the U-shaped serial position effect, the congruity effect and speed-accuracy tradeoffs all together. The testing of all possible combinations is now a standard practice for comparative judgements studies (e.g., Jou & Aldridge, 1999; Jou, 2003).

1.4.2 Grouping effects

Comparative judgements have also been studied in sets of stimuli that are grouped into discrete groups (usually two groups) and found mixed evidence on how organizing the study list in discrete groups would influence the behavioural results (see Pohl, 1990; Jou, 2011, for reviews). Some researchers (Holyoak & Patterson, 1981; Kosslyn, Murphy, Bemesderfer, & Feinstein, 1977) suggest the group label and the continuous-valued magnitude could both be used for comparison, where the group labels are discrete codes and the magnitude information are analogue codes. For instance, for a list ABC DEF, ABC would be labeled as group X, DEF would be labeled as group Y, in addition to the analogue code 123456 ordered by presentation time (For an example, see Figure **??**). This theory led to the prediction that within-group judgements should show a distance effect, whereas between-group judgements

should show no distance effect, and between-group judgements should be faster or as fast as within-group judgements. Some studies have shown violation of this hypothesis that the distance effect could be found in between-group judgements (Howard, 1980; Kosslyn et al., 1977; Maki, 1982; Woocher et al., 1978) and between-group judgements could be longer than within-group judgements (Kosslyn et al., 1977; Maki, 1981, 1981; Woocher et al., 1978). However, studies showed that the distance effect could be attenuated (Maki, 1981) or disappear (Kosslyn et al., 1977; Pohl, 1990; Pliske & Smith, 1979) when the groups were over-learned (Kosslyn et al., 1977), or of pre-existing semantic categories (Howard, 1980; Maki, 1981; Pliske & Smith, 1979; Sailor & Shoben, 1993; Shoben & Wilson, 1998), or when the serial position effect could be controlled (Pohl, 1990). This set of results suggests participants could make judgements by comparing the group labels, but it is a less efficient strategy than comparing the relative magnitude directly, especially when the group information is not already well learned and easily retrievable. The two strategies could compete with each other to maximize efficiency (Cech & Shoben, 2001). Unlike the distance effect, the grouping has not been found to influence the serial position effect. When the data are organized by the position of the lower ranking probe and when probe distance is kept constant, comparative judgements produce an inverted-U shaped serial position curve, showing little difference between grouped and ungrouped lists (Woocher et al., 1978; Jou, 2005, 2011). Jou (2011) made the connection between comparative judgements and serial recall, suggesting the lack of "scalloped" pattern in comparative judgements results is task-specific, that local reference points help serial recall, but has little utility for making comparative judgements. This could explain why empirical data from comparative judgements favour uni-dimensional memory structure, whereas empirical data from serial recall favour hierarchically organized memory structure. However, we need to compare error-rate results to fully understand the grouping effect difference between JOR and serial recall test, which we will do in Chapter 5. As we already discussed, group labels are considered as discrete codes (Holyoak & Patterson, 1981; Kosslyn et al., 1977). In other words, when categories are used as groups, the group labels are considered to be nominal (e.g., Maki, 1981; Sailor & Shoben, 1993). Studies on category effects using comparative judgements usually studies two groups, thus could not test whether the group labels could convey ordinal information. However, positional coding theories suggest group labels are ordinal. For example, Group code 1 is smaller than Group code 2, and Group code 2 is smaller than Group code 3, this would predict a distance

effect for group codes. In Chapter 5, we study the effects of group labels on the congruity effect and serial position effect. Our data show between-group judgements have a distance effect, and larger group code distance enhances judgement accuracy. However, we could not identify meaningful grouping effects on the serial position plots. Although no "scalloped" serial position pattern is found, grouping effects are in line with specific JOR predictions based on two-level hierarchical position codes.

1.5 Summary

To summarize, first, the congruity effect discovered by Chan et al. (2009) generalizes to both subspan and supraspan lists, semantic and episodic memory, grouped and ungrouped lists. Second, this suggests commonalities of the congruity effect in comparative judgements and temporal order judgements. Finally, although developed from different fields of order memory, grouping effects on both the JOR and serial recall results show evidence of hierarchical positional coding. This suggests both tasks may share similar underlying mechanisms.

1.6 Chapters overview

The next chapters in this dissertation follow a logical structure to test the boundary conditions of the JOR congruity effect, to explore how JOR results are related to comparative judgements and serial recall results, and to discuss how JOR results could contribute to advancing order memory theories. Here we briefly outline the key aspects of the following chapters. In Chapter 2, we test whether the subspan JOR congruity effect could be found in supraspan lists, using noun lists with list lengths 4, 6, 8 and 10 and consonant lists with list lengths 4 and 8, and find that the congruity effect not only generalized to supraspan lists, but could also be found using the error rate measure. The congruity effect of subspan lists are better explained by a switching search-direction in Hacker's (1980) self-terminating search model, whereas the congruity effect of supraspan lists is better explained by adding a bias parameter to SIMPLE (Brown et al., 2007). In Chapter 3, we test whether the congruity effect could generalize to very long lists, and we ask participants to judge relative order of the English alphabets. We find a congruity effect using the response time measure, and an error rate congruity effect, masked by speed-accuracy tradeoffs. In Chapter 4, we test the grouping effects on forward and backward serial recall, using a 9-item consonant list

with temporally induced 3 groups of 3 items versus a 9-item consonant list with even inter-stimulus interval, matched on the total presentation time. We replicate previous key results of forward serial recall, and find that although backward recall is overall slower than forward recall, the grouping effects are qualitatively similar between instructions when the data is aligned using output positions. The results are consistent with predictions of a two-level positional coding model if we assume positions could be retrieved directly in the backward order. In Chapter 5, the materials and presentation procedure are identical to Chapter 4. However, we ask participants to perform relative order judgements instead of serial recall. We find that the congruity effect generalizes across grouping conditions. In addition, we find that grouping enhanced JOR accuracy at the cost of increased response times, and the JOR results are consistent with specific predictions based on a two-level positional coding model. In Chapter 6, we summarize the important findings from Chapters 2 to 5 and further discuss the implications of our findings on advancing theories of order memory.

Chapter 2

Generality of a congruity effect in judgements of relative order

2.1 Introduction

In remembering everyday information, such as a telephone number, a route or a sequence of events, order is central (Lashley, 1951). A relatively simple test of memory for order is the judgement of relative order (JOR) procedure (Butters, Kaszniak, Glisky, Eslinger, & Schacter, 1994; Chan et al., 2009; Fozard, 1970; Hacker, 1980; Hockley, 1984; Hurst & Volpe, 1982; Klein, Shiffrin, & Criss, 2007; McElree & Dosher, 1993; Milner, 1971; Muter, 1979; Naveh-Benjamin, 1990; Wolff, 1966; Yntema & Trask, 1963). Illustrated in Figure 2.1, the JOR procedure tests memory for relative order without requiring participants to produce the items from memory. The wording of a JOR question typically takes a form like, "Which of two people left the party more recently?" A logically equivalent form of this question could be: "Which of two people left the party earlier?" Because formally, all that has changed is that the target became the non-target and vice-versa, one might presume that these "earlier" and "later" instructions test the same information in memory. Perhaps this is why few studies have compared these instructions. The vast majority have used a "recency" instruction, hence the term, "judgement of relative recency" (the origin of the acronym, JOR). However, instructions do influence JOR performance on both supra- and

A version of this work was previously published as: Liu, Y. S., Chan, M., & Caplan, J. B. (2014), "Generality of a congruity effect in judgements of relative order". *Memory & Cognition , 42* (7), 1086-1105.

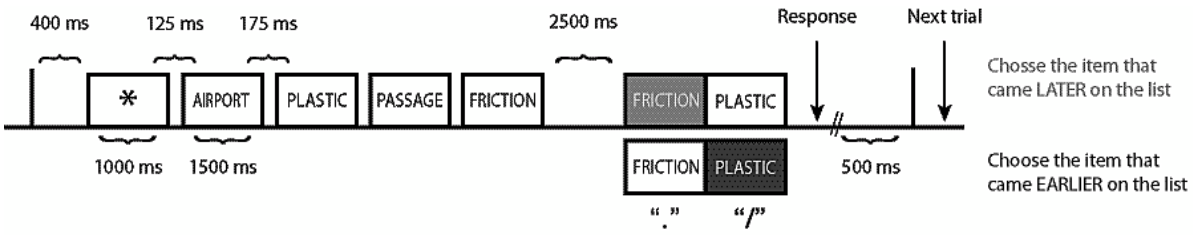


Figure 2.1: Time course of one example experimental trial in Experiment 1 (list length=4 nouns) with both instructions. At test, two nouns from the list are presented in random order, and the participant is asked to respond to the probe stimulus that occurred earlier ("earlier" instruction) or later ("later" instruction) in the just-presented list. The correct response item is depicted on a dark background in this figure only, not in the experiment itself. The keyboard key that the participant would press to select each probe item is depicted underneath the probe items.

sub-span lists: Flexser and Bower (1974) found that their "distant" instruction had worse overall accuracy than their "recency" instruction. More specifically, Chan et al. (2009) found that participants' behaviour on sub-span lists resembled backward, self-terminating search for a "later" instruction, consistent with previous findings (Hacker, 1980; Muter, 1979), but *forward,* self-terminating search for an "earlier" instruction. Here we ask whether this congruity effect is confined to sub-span lists, or generalizes to longer, supra-span lists.

Figure 2.2c illustrates how hypothetical response-time data would look for a forward, self-terminating search strategy. The vertical axis plots the behavioural measure; for illustration purposes we label it "error rate" or "response time," because speed–accuracy tradeoffs notwithstanding (and we found none in our data), one would expect response time and error rates to vary in the same direction as one another. The left horizontal axis plots the serial position of the earlier probe item, and the right horizontal axis plots the serial position of the later probe item. Note that the later-item serial position is plotted in descending order to minimize the bars occluding one another. In forward, self-terminating search, response time/error rate increases as a function of the earlier probe serial position, whereas the later probe serial position has no influence on response time/error rate. The opposite pattern is expected for backward self-terminating search, where response time/error rate increases when the later probe serial position decreases (Figure 2.2d). The effect of instruction can be most clearly visualized if we plot the difference between "earlier" and "later" instruction data (Figure 2.2e).

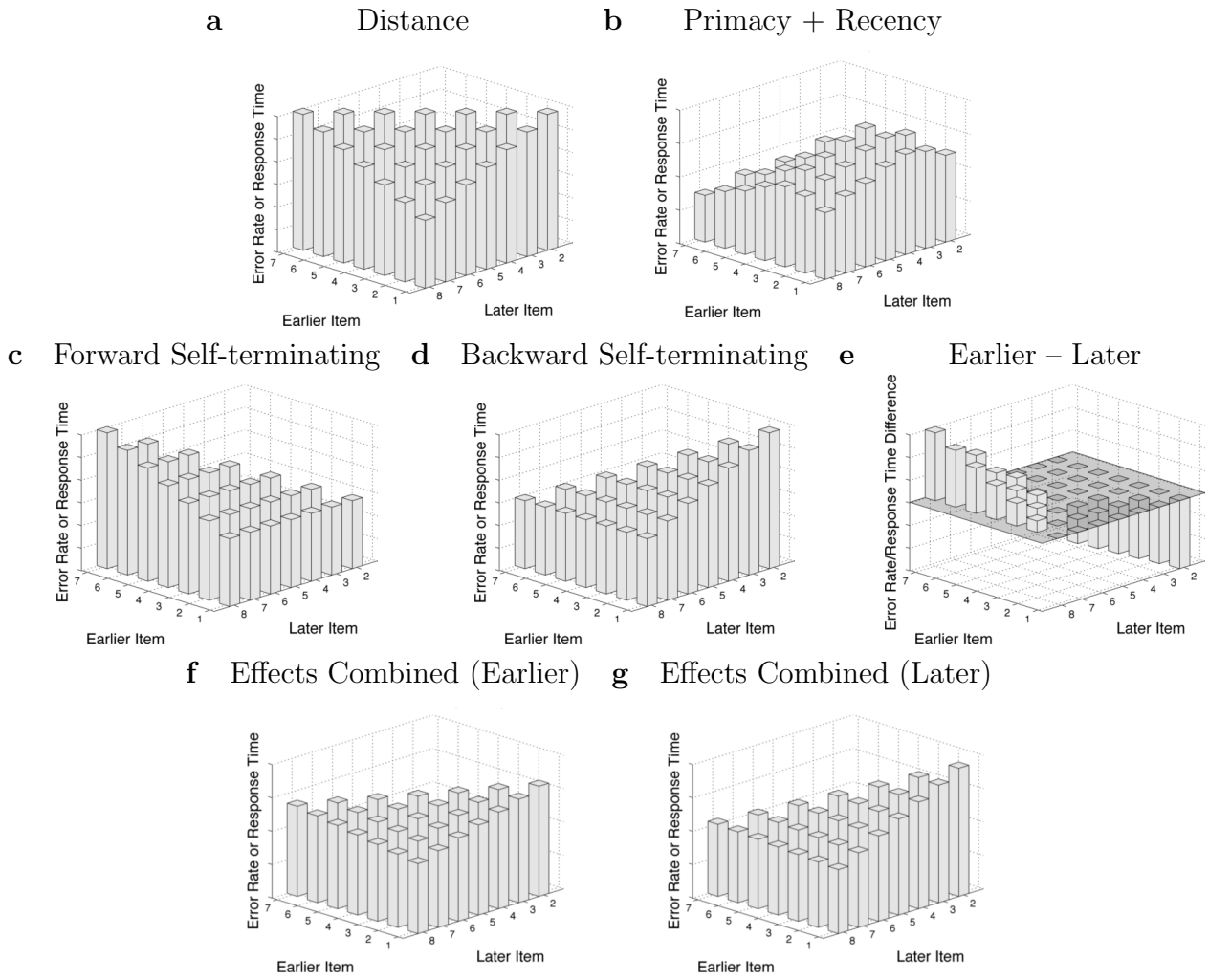


Figure 2.2: Schematic depictions of hypothesized serial position effects. The dependent measure (error rate or response time) is plotted as a function of both the earlier probe-item's serial position ("Earlier Item") and later probe-item's serial position ("Later Item"). **a,** Serial position effects expected due to the distance effect. **b,** Serial position effects expected due to the primacy and recency effect. **c,** Serial position effects for forward, self-terminating search, as was found in sub-span lists using "earlier" instruction (Chan et al., 2009). **d,** Serial position effects for backward, self-terminating search, as was found in sub-span lists using "later" instruction (Chan et al., 2009). **e,** The difference between (a) and (b), which we use to isolate the congruity effect. **f,** Our hypothesized serial position effects for "earlier" instruction for supra-span lists: an average of recency, distance and instruction-based bias across the list. **g,** Our hypothesized serial position effects for "later" instruction, as an average of recency, distance and instruction-based bias across the list. Note that the hypothesis for the difference between instructions for supra-span lists remains as in (e), except that edge effects are expected to produce bow-shaped, rather than linear congruity effects.

We already know that JORs for supra-span lists are qualitatively quite different, and two important findings may suggest we would not find a congruity effect at longer list lengths: (a) a distance effect (Figure 2.2a), whereby judgements are better (faster and more accurate) as the difference in serial positions (distance) of the two probe-items increases (e.g., Bower, 1971; Yntema & Trask, 1963), similar to the symbolic distance effect (e.g., Banks, 1977; Holyoak, 1977; Moyer & Landauer, 1967); and (b) an inverted U-shaped serial position effect, comprised of a primacy and recency effect (Figure 2.2b) (e.g., Hacker, 1980; Jou, 2003; Muter, 1979; Yntema & Trask, 1963). Chan et al.'s congruity effect was found for response times, suggesting that instruction influenced access-speed as a function of serial position. For supra-span JORs, error rate is also a useful dependent measure. As list length increases above span, error rate increases; in an extreme case, with a list length of 90 words, accuracy approached chance-levels, rising to 60% accuracy only for very large lags (distance of 36 words; Klein et al., 2007). Primacy and recency effects may seem at odds with self-terminating search models that are reasonable accounts of sub-span data (Chan et al., 2009). However, Hacker (1980) suggested that, in the case of imperfect item-memory, U-shaped serial position effects due to item-memory might distort self-terminating search patterns in JORs, an idea he incorporated into his self-terminating search model. The distance effect is also incompatible with self-terminating search, because the position of the unreached probe item should not affect the outcome of the JOR decision. These arguments might lead one to expect no congruity effect in long lists.

On the other hand, there are reasons to expect there should be a congruity effect at long list lengths. Evidence suggests there is no clear distinction between short- and long-term order-memory (McElree, 2006). Moreover, Muter (1979) found a backward self-terminating search pattern extending to lists of ten items (supra-span). Hacker's (1980) data did not show obvious break points of his "availability" parameter (representing item-memory) that could have distinguished a working memory from a long-term memory. This is consistent with extensive evidence suggesting that memory is scale-invariant (Brown et al., 2007; Crowder, 1982; Howard & Kahana, 1999; Nairne, 2002). We suggest it is possible both long and short list lengths are governed by the same memory mechanisms, and the congruity effect will generalize from short to longer list lengths.

In addition, the self-terminating search model has been fitted to long-list JOR data with success (Hacker, 1980; McElree & Dosher, 1993). It is possible that a self-terminating search

model operating in the forward, rather than the backward, direction could explain "earlier" instruction data and thus account for the congruity effect. Thus, "earlier" instruction might induce a dominant primacy effect even for longer lists. In serial-recall procedures, forward recall shows a dominant primacy effect, whereas backward recall shows a dominant recency effect (Beaman, 2002; Hulme et al., 1997; Li & Lewandowsky, 1993, 1995; Li et al., 2010; Madigan, 1971; Richardson, 2007; Rosen & Engle, 1997; Thomas et al., 2003), suggesting that if forward-search is based on serial recall, this kind of mechanism might be applicable even for longer lists. At present, published studies of supra-span JORs have mainly used a "recency" instruction to look at serial position effects, similar to our "later" instruction (Butters et al., 1994; Chan et al., 2009; Fozard, 1970; Hacker, 1980; Hockley, 1984; Hurst & Volpe, 1982; Klein et al., 2007; McElree & Dosher, 1993; Milner, 1971; Muter, 1979; Naveh-Benjamin, 1990; Wolff, 1966; Yntema & Trask, 1963). Wyer, Shoben, Fuhrman, and Bodenhausen (1985) used both "sooner" and "later" instructions with probes derived from a social-action script (e.g., going to a restaurant), and found a response time congruity effect, but not for events that were specific to the example story. A similar response time congruity effect was found for personal life events in a subset of experimental conditions (Fuhrman & Wyer, 1988). These congruity effects for action scripts and personal life events may reflect supra-span phenomena, but both types of material are arguably tapping into semantic, not episodic, temporal order. We wondered if the JOR-congruity effect would generalize above span, with response time as the measure.

Since we expected error rate to be an informative dependent measure for these lists, we wondered whether instruction would affect the quality of information in memory (availability), measured by error rate, or just accessibility, measured by response time. An error rate congruity effect has been found in autobiographical order tasks with yes/no judgements (Skowronski, Walker, & Betz, 2003; Skowronski et al., 2007); however, participants' confirmation bias (toward selecting "yes" rather than "no") might underlie that result. We found no clear published error rate congruity effect for temporal-order memory, although error-rate congruity effects have occasionally been found for perceptual comparative judgements (Petrusic, 1992). We therefore hypothesized that a similar congruity effect would be observed in supra-span JOR data, but with the addition of recency, primacy and distance effects, with both response time and error rate as measures. If we assume that the primacy, recency and distance effects are approximately constant between instructions, we can isolate

the congruity effect by analyzing the difference between instructions (Figure 2.2e), which would then look similar to that observed in sub-span response time data (Chan et al., 2009). We test these hypotheses in two experiments, always manipulating instruction between subjects. Experiment 1 used lists of nouns, and manipulated list length (4, 6, 8 and 10) within subjects. Experiment 2 used consonant lists, and manipulated list length (4 and 8) between subjects. The experiments produced similar results, suggesting broad boundary conditions for the congruity effect. Experiment 2 used the same materials and presentation rate as Chan et al.'s (2009) experiment.

To broaden the theoretical implications of our results, we evaluated our findings with respect to Hacker's (1980) self-terminating search model. Hacker developed this model specifically to explain JORs, but it has not been tested on the congruity effect. We hypothesize the congruity effect can be explained by a difference in the direction of search associated with each instruction. Participants may perform forward, self-terminating search with "earlier" instruction, and backward, self-terminating search with "later" instruction, and we test this with fits of models based on Hacker's model after presenting the results of both experiments. We also discuss whether other existing memory models for JOR paradigm could account for the congruity effect in their current form, or could be easily adapted to do so.

2.2 Experiment 1

2.2.1 Methods

Participants

Fourteen participants were recruited from the University of Alberta community. Participants gave informed consent and were paid at a rate of $12 for each of five 1-h sessions, conducted on five consecutive days. All had normal or corrected-to-normal vision and had learned English before the age of 6. Participants were randomly assigned to "earlier" or "later" group in alternating testing order. One participant in "later" instruction did not attend the last session, so for that participant, only the first four sessions were included in the analyses.

Materials

Stimuli were 1316 nouns generated from the MRC Psycholinguistic Database (Wilson, 1988) with word length restricted to three to eight letters, two syllables and Kucera-Francis written

frequency above 6 per million, displayed in uppercase. Nouns that we subjectively determined might be confused with verbs were manually removed from the list. Each trial was randomly drawn from list length 4, 6, 8, and 10, counterbalanced within-session. There was no within-session repetition of words, but words were re-used across sessions. All participants were tested using an A1207 iMac computer with an Apple Macintosh A1048 Pro keyboard.

Procedure

The experiment was implemented with the Python Experiment-Programming Library (PyEPL; Geller, Schleifer, Sederberg, Jacobs, & Kahana, 2007) and modified from Chan et al.'s (2009) experiment (Figure 2.1). Probes were pairs of items drawn from the just-presented list, and all possible combinations were equally probable and counterbalanced within subject and within list length. Participants in the two groups received slightly different instructions: (a) Excerpt from "earlier" instruction: "...judge which of the two nouns came earlier on the list you just studied. Press the '/' key if the earlier item is presented on the right side of the screen and the '.' key if the earlier item is on the left side of the screen. ..." (b) Excerpt from "later" instruction: "...judge which of the two nouns came later on the list you just studied. Press the '/' key if the later item is presented on the right side of the screen and the '.' key if the later item is on the left side of the screen...". Participants were instructed to respond as quickly as they could without compromising accuracy. A session consisted of 9 blocks with 20 trials in each block. The first block of each session was a practice block, excluded from analyses, composed of 8 trials, to familiarize (or re-familiarize) participants with the task. The computer provided immediate accuracy feedback after each trial in practice block ("correct" or "incorrect"), and average response time (ms) and accuracy (%correct) at the end of each experimental block. Each trial began with a fixation asterisk, '*', in the center of the screen, followed by a word list presented sequentially in the center of the screen. Items were presented for 1500 ms each with an inter-stimulus interval (ISI) of 175 ms. This is slower than the rate Chan et al. (2009) used (575 ms presentation time and 175-ms ISI), due to the greater stimulus complexity of nouns compared to consonants (e.g., Sternberg, 1975). After a 2500-ms delay, participants were presented with a single probe consisting of two words from the just-presented list and were asked which item was presented earlier or later, depending on group, by pressing '.' key (for the left-hand probe item) or the '/' key (for the right-hand probe item). After a 500-ms delay, participants could press a key to start

the next trial.

Data analysis

Trials with response time less than 200 ms and above three standard deviations from a participant's mean response time were removed from the data (1.3% of responses). A linear mixed effects (LME) model (Baayen, Davidson, & Bates, 2008; Bates, 2005) was used to analysis our data. We adopted LME analysis because compared to ANOVA, LME handles unbalanced designs, can fit individual responses without the need for averaging of the data, and protects against type II error due to increased power (Baayen et al., 2008; Baayen & Milin, 2010). LME analyses were conducted in R (Bates, 2005), using the LME4 (Bates & Sarkar, 2007), LanguageR (Baayen, 2007) and LMERConvenienceFunctions (Tremblay, 2013) libraries. The "lmer" function was used to fit the LME model. The "pamer.fnc" function was used to calculate the p values of model parameters. Eight fixed factors were used as predictors, including Instruction ("earlier", "later'), linear and quadratic component of Later-Probe Serial Position (serial position of the probe item that appeared later from the presented list), Distance (absolute value of the difference between two probe's serial positions), Intact/Reverse (whether probe order was consistent or inconsistent, with presentation order, respectively), Trial Number, Session Number, and List Length. The linear and quadratic component of the Later-Probe Serial Position are orthogonal to each other, generated with the "poly" function in R. We included the quadratic term to account for expected primacy and recency effect. Subject was included as a random effect on intercept. Instruction and Intact/Reverse were treated as categorical factors. All other factors were scaled and centered before being entered in the model. Response time was analyzed for correct trials only, and was log-transformed to reduce skewness. The error rate data were fitted with logistic regression as it is a binary variable ("correct" vs. "incorrect"). LME estimated random effects first, followed by fixed effects. In the results tables, the "Estimate" column reported the corresponding regression coefficients, along with their standard errors. For the purposes of reporting the LME results, the Intact condition and "earlier" instruction were set as the reference levels for the Intact/Reverse and Instruction factors, respectively. The best fits of LME models were obtained by conducting a series of iterative tests comparing progressively simpler models with more complex models using the Bayesian Information Criterion (BIC). We used BIC because it penalizes free parameters more than

the Akaike Information Criterion (AIC), making it conservative and resistant to over-fitting (Motulsky & Christopoulos, 2004; Zuur, Leno, Walker, Saveliev, & Smith, 2009). This approach is adopted to remove interactions and variables that do not explain significant amount of variance (Baayen et al., 2008). We used LMERConvenienceFunctions (Tremblay, 2013) library to conduct fitting of fixed effects systematically. In this approach, for each condition we started with a model that included all factor combinations and interactions with two exceptions: a) The quadratic component of Later-Probe Serial Position was not allowed to interact with the linear component of Later-Probe Serial Position because both were derived from the Later-Probe Serial Position. b) Any interaction term for which one or more levels had no data. Starting with the complete model, the highest-order terms are considered first, progressing to the lowest-order terms. At each stage, considering a given order of interaction, the term with the lowest p value is identified and a model without this term is compared with the original model using BIC. The term is kept if it improves BIC based on a threshold of 2 or if the term is also contained within a higher-order interaction. When all terms are tested for the highest-order interaction, the comparison process continues to the term with lowest p value in the next highest-order interaction, and so on. The process iterates until all interaction terms have been tested, ending with main effects (Tremblay, 2013).

2.2.2 Results and Discussion

Error rate and response time, averaged across participants, are plotted as functions of serial position of the earlier and later probe items in Figures 2.3 and 2.4. We isolated the congruity effect by plotting the difference between "earlier" and "later" instructions after first removing the overall mean for each participant (right-hand columns). The best-fitting LME model is reported in Table 2.1 and 2.2. To better visualize the pattern of serial-position effects, the overall mean was removed to correct for the mean difference between "earlier" and "later" instruction.

Error rates

First, we replicated the known effects of bow-shaped serial position effects and distance effects. At all list lengths and for both instructions, the error rate data (Figure 2.3) showed a distance effect (Figure 2.2a), supported by a significant main effect of Distance, and bow-shaped serial position effect involving both primacy and recency (Figure 2.2b), supported by

	Estimate (SE)
—Main effects—	
Intercept	-2.99 (0.29)*
Intact/Reverse	0.531 (0.090)*
Later-Probe Serial Position (Quadratic)	-51.4 (5.7)*
Instruction	0.61 (0.39)
Distance	-0.612 (0.061)*
Trial	-0.082 (0.032)*
List Length	1.225 (0.055)*
Session	0.086 (0.032)*
Later-Probe Serial Position (Linear)	-79.8 (8.0)*
—Interactions—	
Intact/Reverse × Instruction	-1.32 (0.13)*
Trial × Session	0.122 (0.032)*
Instruction × Later-Probe Serial Position (Linear)	-35.4 (6.9)*
Distance × Later-Probe Serial Position (Linear)	26.5 (5.3)*
LL × Later-Probe Serial Position (Linear)	42.4 (6.0)*

Table 2.1: The best-fitting LME model for experiment 1 error rate results. The congruity effect is in bold. The "Estimate" column reports the corresponding regression coefficient, along with its SE (standard error). Significant effects are denoted * - $p < 0.05$.

significant quadratic component of the Later-Probe Serial Position in the best-fitting LME model (Table 2.1). The "later" instruction (Figure 2.3, middle column) broadly resembled "earlier" instruction (Figure 2.3, left-hand column) except that the recency effect was more pronounced for "later" instruction.

We next asked whether, despite the presence of distance and serial-position effects, there might also be a congruity effect. The difference bar graph (Figure 2.3, right-hand column) shows that instruction indeed interacted with probe serial positions, supported in the LME analysis by interactions between Instruction and linear component of Later-Probe Serial Position (Table 2.1). This interaction was due to "earlier" instruction producing better performance at earlier serial positions, and "later" instruction producing better performance at later serial positions, in line with our predicted congruity effect (Figure 2.2e).

Additional findings of interest that emerged from the best-fitting LME model were main effects of List Length, Intact/Reverse, Trial and Session. More error was associated with greater list length, reverse probe presentation order, lower trial number and lower session number.

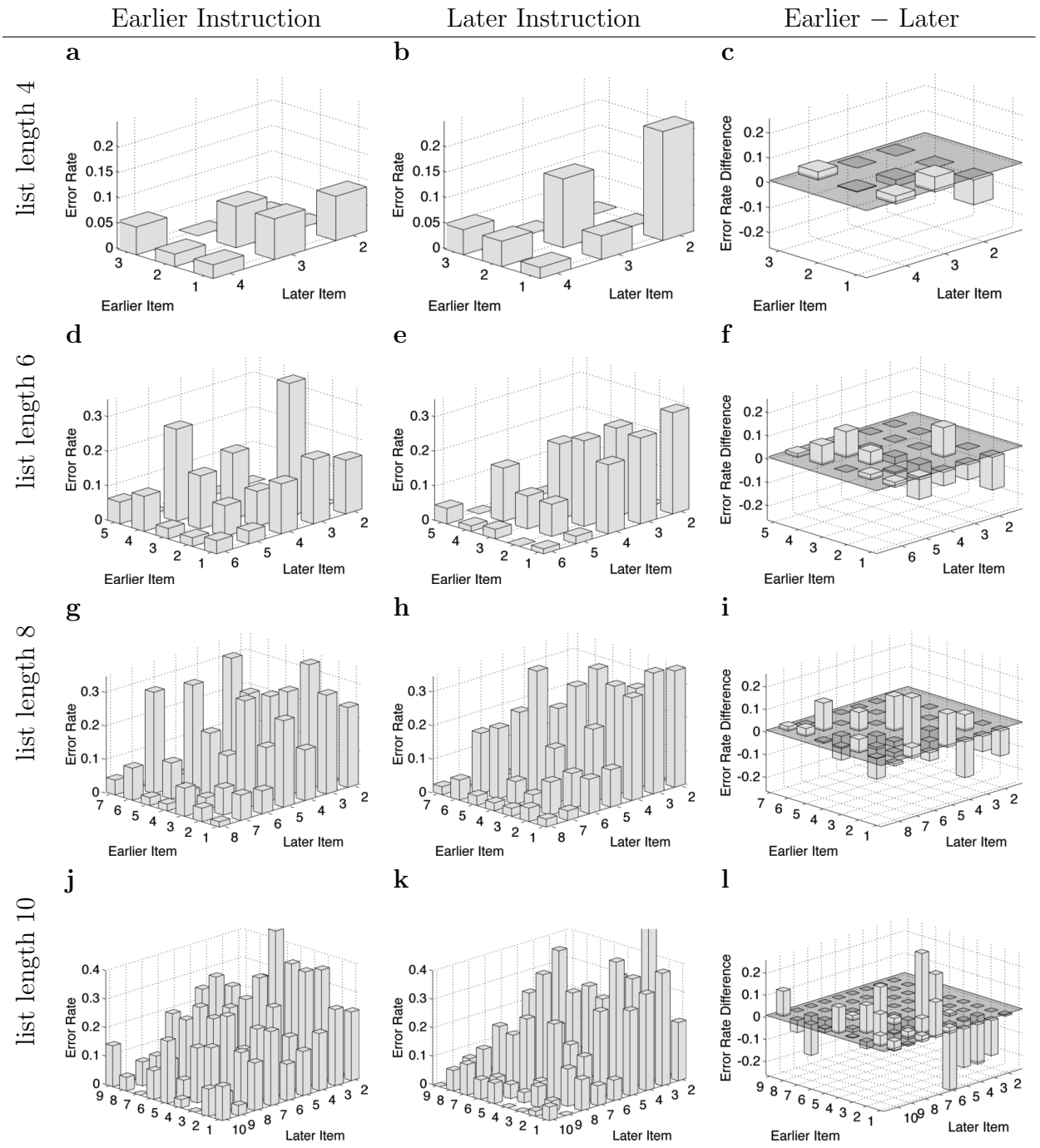


Figure 2.3: Error rate (Experiment 1) as a function of both probe items' serial position (earlier item and later item, respectively) broken down by list length in rows, and instruction ("earlier", "later" and the difference, "earlier"−"later", corrected for mean error rate) in columns.

Importantly, list length did not interact with the congruity effect, suggesting the congruity effect on error rate is replicated at all list lengths and does not change substantially across our four list lengths. We found a significant Trial × Session interaction. The interaction is consistent with learning-to-learn effects; larger trial numbers have less errors, and this effect reduces in later sessions. Importantly, Trial and Session both did not interact with the congruity effect, showing that the congruity effect generalizes across these factors.

Finally, a significant interaction was found for Instruction × Intact/Reverse. This is a second kind of congruity effect between instruction and reading order: Intact probes were judged better for "earlier" instruction and worse for "later" instruction. Reverse probes had the opposite relationship to instruction. If participants read from left to right, this would indicate better performance when the target was read first.

Response times

First, as with error rate, for all list lengths and both instructions, the response time data (Figure 2.4) had significant distance and bow-shaped serial position effects (Figure 2.2a), supported by a significant main effect of Distance and quadratic component of Later-Probe Serial Position, respectively, in the best-fitting LME model (Table 2.2).

Turning to the congruity effect, as with error rate, the difference bar graph (Figure 2.4, right-hand columns) shows the predicted congruity effect, supported in the LME analysis by significant interactions between Instruction and linear component of Later-Probe Serial Position (Table 2.2). Again, in line with our predicted congruity effect (Figure 2.2e), "earlier" instruction produced better performance at earlier serial positions, and vice versa for "later" instruction.

We further checked whether the congruity effect was qualified by significant three-way interactions in the best-fitting LME model. The three-way interaction of Instruction × linear component of Later-Probe Serial Position × Distance showed increasing Distance was associated with a decrease in the slope of the linear component of Later-Probe Serial Position for both instructions (see Figure S1 in supplementary materials). However, the rate of the linear component of Later-Probe Serial Position function's slope decrease was steeper for "earlier" instruction than for "later" instruction. The differential rate of slope decrease, thus, does not contradict the congruity effect. The interaction of Instruction × quadratic component of Later-Probe Serial Position × List Length showed a pattern of decreasing quadratic com-

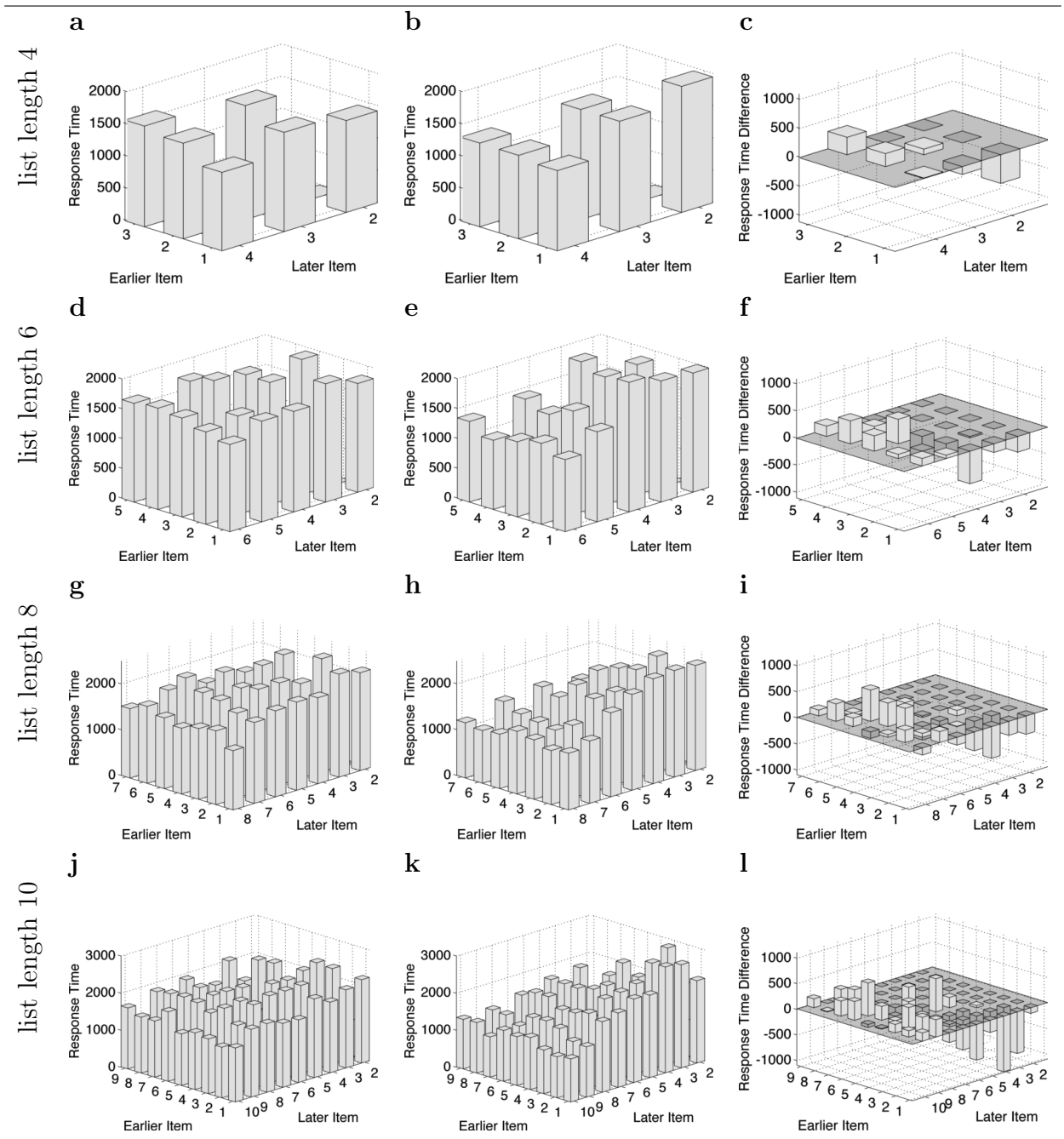


Figure 2.4: Response time (Experiment 1) as a function of both probe items' serial position (earlier item and later item, respectively) broken down by list length in rows, and instruction ("earlier", "later" and the difference, "earlier"−"later", corrected for mean response time) in columns.

	Estimate (SE)
—Main effects—	
Intercept	7.240 (0.072)*
LL	0.277 (0.011)*
Instruction	-0.01 (0.10)
Intact/Reverse	0.038 (0.013)
Trial	-0.0147 (0.0063)*
Distance	-0.125 (0.012)*
Session	-0.141 (0.006)*
Later-Probe Serial Position (Linear)	-17.5 (1.5)*
Later-Probe Serial Position (Quadratic)	-18.1 (1.3)*
—Interactions—	
LL × Instruction	0.056 (0.013)*
LL × Distance	-0.015 (0.011)
LL × Session	0.0300 (0.0063)*
LL × Later-Probe Serial Position (Linear)	10.7 (1.2)*
Instruction × Intact/Reverse	-0.082(0.018)*
Instruction × Trial	-0.03217 (0.0089)*
Instruction × Distance	0.089 (0.016)*
Instruction × Session	-0.0842 (0.0089)*
Instruction × Later-Probe Serial Position (Linear)	-13.7 (1.7)*
Trial × Session	0.0190 (0.0045)*
Distance × Later-Probe Serial Position (Linear)	7.7 (1.2)
Session × Later-Probe Serial Position (Linear)	-2.65 (0.62)*
LL × Later-Probe Serial Position (Quadratic)	3.25 (0.76)*
Instruction × Later-Probe Serial Position (Quadratic)	10.9 (1.3)*
Instruction × Distance × Later-Probe Serial Position (Linear)	-6.1 (1.2)*
LL × Instruction × Later-Probe Serial Position (Quadratic)	-6.3 (1.0)*

Table 2.2: The best-fitting LME model for experiment 1 response time. The congruity effect is in bold. The "Estimate" column reports the corresponding regression coefficient, along with its SE (standard error). Significant effects are denoted * - $p < 0.05$.

ponent of Later-Probe Serial Position slope for "later" Instructions and increasing quadratic component of Later-Probe Serial Position slope for "earlier" Instruction, as List Length increases (see Figure S2 in supplementary materials). This interaction suggests the difference in the primacy and recency effects between instructions decreases as list length increases.

Similar to the error rate results, we found Trial × Session and Instruction × Intact/Reverse interactions. Instruction also interacted with Trial, Session, and Distance. Response time in "later" instruction improved more with practice than the in "earlier" instruction. The "later" instruction also had a smaller distance effect than "earlier" instruction. List Length interacted with Instruction, Session and Later-Probe Serial Position. To summarize this effect, increasing list length was associated with slower response times for "later" instruction, higher session number and larger Later-Probe Serial Position.

In sum, experiment 1 replicated the typical primacy, recency and distance effects (Hacker, 1980; Jou, 2003; Muter, 1979; Yntema & Trask, 1963), and extended Chan et al.'s (2009) congruity effect finding from sub-span (e.g., list length 4) to supra-span data (up to list length 10). The congruity effect appeared in both error rate and response time measures.

2.3 Experiment 2

One potential confound in experiment 1 is that participants were given four list lengths, intermixed. It is possible that that the congruity effect is in fact a sub-span— not supra-span— phenomenon, but that the inclusion of some sub-span lists (list length 4) influenced participants to apply a sub-span strategy to supra-span lists. Thus, perhaps our congruity effect in supra-span lists is a special case. To address this, list length was a between-subjects factor in experiment 2. In addition, to test for boundary conditions of the congruity effect, we switched from nouns to consonants and to a faster presentation rate (similar to the one used by Chan et al., 2009). If the congruity effect were found regardless of practice effects, stimulus type and presentation rate, the generality of congruity effect would be further supported.

2.3.1 Methods

Participants

A total of 385 undergraduate students from introductory psychology courses at the University of Alberta participated in exchange for partial course credit. Participants gave informed consent, had normal or corrected-to-normal vision and learned English before age 6. We included two between-subjects factors: list length (4, 8) × Instruction ("earlier", "later"). Participants were run in groups of about 10–15 with all participants within a testing group being assigned to a single experimental group; experimental group cycled across testing groups. Forty-four participants were excluded because their error rate was close to chance ($\geq$ 40%). The number of excluded versus included participants in each condition is summarized in Table 2.5.

Materials

Materials were the same as those used by Chan et al. (2009). Stimuli were 16 consonants (excluding S, W, X, and Z) from the English alphabet displayed in uppercase. Each list comprised 4 or 8 (depending on group) consonants drawn at random without replacement from the stimulus pool, with the restriction that they did not appear in the two preceding lists. Probability was equal for each consonant/serial-position combination. All participants were tested using a group of 15 computers (custom-built PCs) with identical hardware, identical Samsung SyncMaster B2440 monitors and Logitech K200 keyboards. Therefore both instruction groups were exposed to the same hardware precision variabilities (Plant & Turner, 2009); thus we do not expect any bias in our between-subjects design.

Procedure

The experiment was again created and run using the Python Experiment-Programming Library (Geller et al., 2007). A single session lasted approximately one hour. The session started with a practice block of 8 trials, followed by 9 blocks of 20 trials each for list length 4, and 6 blocks of 20 trials each for list length 8. The different number of blocks ensured that all participants could finish within one hour. The computer provided online correctness feedback after each trial in practice block ("correct" or "incorrect"), and average response time (ms) and accuracy (%correct) at the end of each block. The instructions were the same

	BIC	AIC	Log-likelihood	df
Best fitting LME model + Congruity effect	10119	19048	-9515.0	9
Best fitting LME model	10120	19057	-9520.6	8
Model difference	-1	-9	5.6	1

Table 2.3: Model comparison of best fitting LME model minimizing BIC to the same model plus Instruction × linear component of Later-Probe Serial Position. Note that for BIC and AIC, lower numbers indicate better fit but for log-likelihood, higher numbers indicate better fit. The log-likelihood ratio test using χ^2 test was significant ($\chi^2 = 11,\ p < 0.05$).

as experiment 1 except the word "nouns" was replaced with "consonants." For each trial participants were first presented with a fixation asterisk, '*', in the center of the screen, then followed by a consonant list that was presented sequentially on the center of the screen with list items presented for 575 ms each with an ISI of 175 ms. After a 2500-ms delay, participants were presented with a two-item probe that consisted of two consonants from the just-presented list and were asked which item was presented earlier/later in the list by pressing the '.' (for the left-hand item) or '/' key (for the right-hand item). Each response was followed by a 500-ms delay before participants could press any key to start the next trial.

Data Analysis

Trials with response time less than 200 ms and above three standard deviations from a participant's mean response time were removed from the data (1.35% of all trials). We adopt the same data representation as in experiment 1. Error rate and response time (correct trials) data were analyzed at each list length separately.

2.3.2 Results and Discussion

Error rates

First, because performance was near ceiling, we could not analyze error rates at list length 4 (Figure 2.5, top row) in any meaningful way. Out of 171 participants for both list length 4 "earlier" and "later" instruction, 89 participants had overall accuracy greater than 95% and only 18 participants scored below 90%. We restrict our error-rate analyses to list length 8 only.

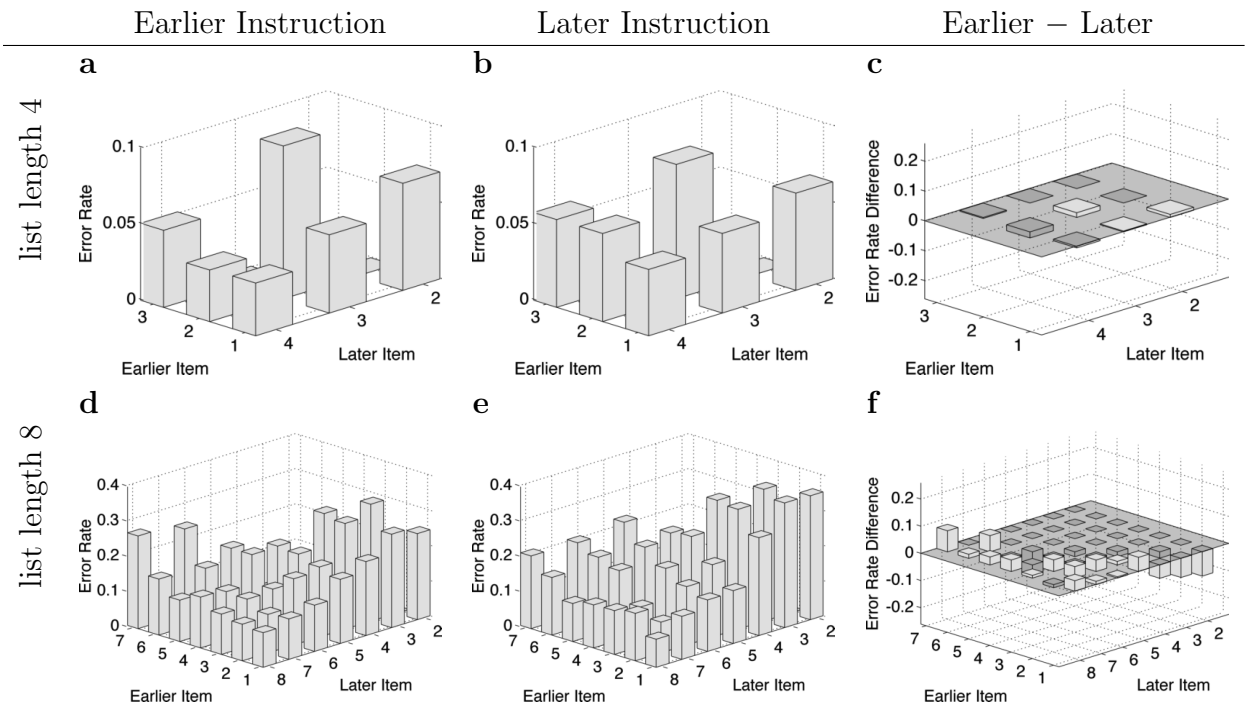


Figure 2.5: Error rate (Experiment 2) as a function of both probe items' serial position (earlier item and later item, respectively) broken down by list length in rows, and instruction ("earlier", "later" and the difference, "earlier"−"later", corrected for mean error rate) in columns.

	Estimate (SE)
—Main effects—	
Intercept	-1.51 (0.06)*
Intact/Reverse	0.41 (0.05)*
Later-Probe Serial Position (Linear)	-40.9 (6.2)*
Instruction	0.74 (0.09)*
Distance	-0.27 (0.02) *
Trial	-0.13 (0.03)*
—Interactions—	
Intact/Reverse × Instruction	-1.04 (0.07)*
Instruction × Later-Probe Serial Position (Linear)	-27.3 (8.2)*

Table 2.4: The best-fitting LME model for experiment 2 list length 8 error rates. The congruity effect is in bold. The "Estimate" column reports the corresponding regression coefficient, along with its SE (standard error). Significant effects are denoted * - $p < 0.05$.

The list-length-8 data (Figure 2.5, bottom row) showed a congruity effect consistent with the pattern observed in experiment 1 (Figure 2.2e), with "earlier" instruction resulting in more errors than "later" instruction as Later-Probe Serial Position increased, supported by a significant Instruction × LateLater-Prober Probe Serial Position (linear component) interaction in the best-fitting LME model (Table 2.4). For this LME model-selection, based on the BIC values, we cannot differentiate the best fitting model with lowest BIC value that included Instruction × Later-Probe Serial Position (the congruity effect), and the same model without the congruity effect term, because $\Delta BIC < 2$. However, because the model that included the congruity effect was *nominally* better by the BIC, we further compared the two models using other fitness criteria. The model that included the congruity effect was reliably selected based on both AIC and log-likelihood (Table 2.3). For this reason, we report the model including the congruity effect. Importantly, the congruity effect did not interact significantly with Trial, Distance or Intact/Reverse, suggesting that it generalizes across these factors.

One can observe an overall recency effect at both list lengths (Figure 2.5), supported by significant Later-Probe Serial Position main effect in the LME model, showing that error rate decreased as Later-Probe Serial Position increased. The distance effect (Figure 2.2a) was also found, supported by a significant main effect of Distance in the best-fitting LME model. There was also a significant main effect of Intact/Reverse and of Instruction; intact probes were better judged than the reverse probes, again suggesting a reading-order effect.

	Earlier list length 4	Later list length 4	Earlier list length 8	Later list length 8
Error rate $\geq 40\%$	2	11	12	19
Total	92	92	99	102

Table 2.5: The number of participants rejected for analysis (error rate $\geq 40\%$) versus total number of subjects in each condition. A chi-square test found differences between number of included subjects for list length 4 and list length 8 were both significant (χ^2=41.2, df=1, $p < 0.001$ and χ^2=4.05, df=1, $p < 0.05$ respectively).

Probes in "earlier" instruction were better judged than in "later" instruction. This is despite more poor performers having been excluded for "later" instruction (Table 2.5); thus, this indicates an overall advantage of "earlier" instruction over "later" instruction. Replicating experiment 1, the Intact/Reverse $\times$ Instruction congruity effect was also significant; intact probes were judged better for "earlier" instruction and worse for "later" instruction. Reverse probes had the opposite relationship to instruction.

Response Time

First, as with experiment 1 error rates and response time results, visual inspection of list length 4 "earlier" instruction found a pattern consistent with forward self-terminating search (Figure 2.2c), and list length 4 "later" instruction found pattern consistent with backward self-terminating search, in line with Chan et al.'s (2009) results. For list length 8, "earlier" instruction pattern resembled a distance effect with an overall primacy and recency effect (Figure 2.2f). The "later" instruction resembled a backward self-terminating pattern combined with distance, primacy and recency effects (Figure 2.2g). The distance effect, primacy and recency effects for both list lengths are supported by a significant main effect of Distance and quadratic component of Later-Probe Serial Position, respectively, in the best-fitting LME models.

Again, replicating the experiment 1 results, the response time data for both list length 4 and list length 8 (Figure 2.6) showed a congruity effect (Figure 2.2e). The congruity effect is supported in the best-fitting model by a significant interaction of Instruction $\times$ Later-Probe Serial Position (linear component) (Table 2.6). The two-way interaction is qualified by a significant four-way interaction of List Length $\times$ Instruction $\times$ Later-Probe Serial Position $\times$ Intact/Reverse. We conducted additional analyses on 4 subgroups of the data: list length

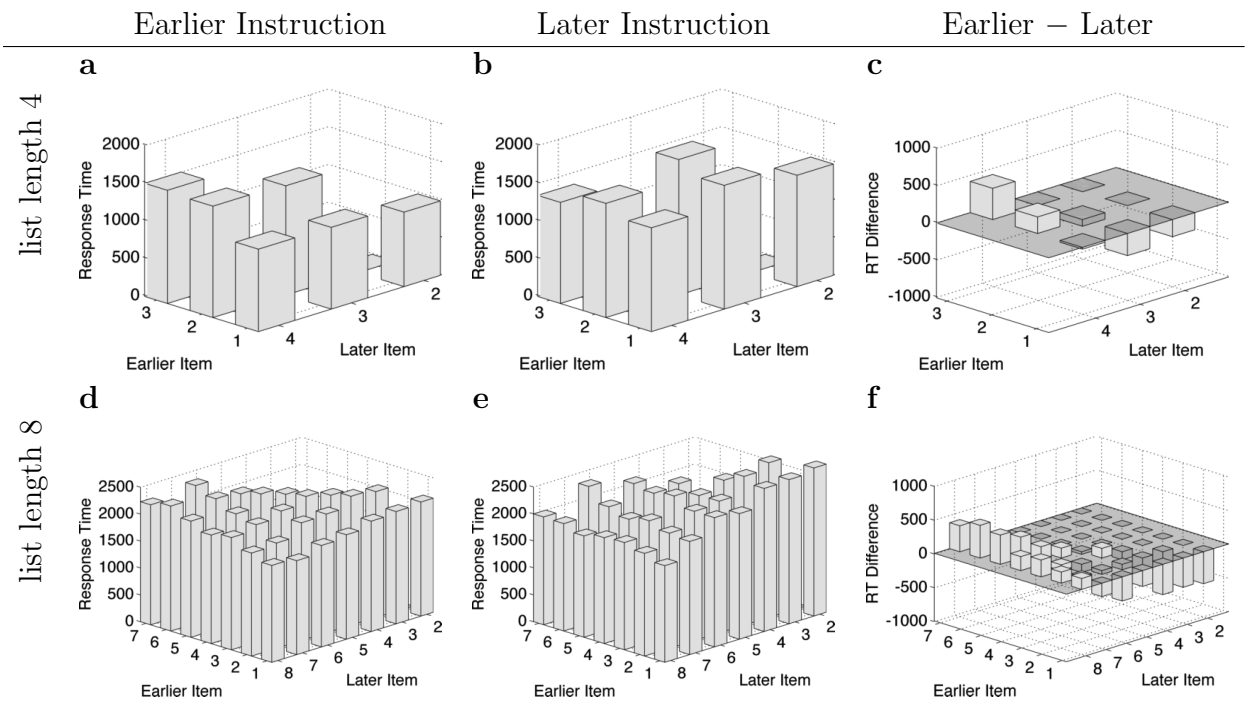


Figure 2.6: Response time (Experiment 2) as a function of both probe items' serial position (earlier item and later item, respectively) broken down by list length in rows, and instruction ("earlier", "later" and the difference, "earlier"−"later", corrected for mean response time) in columns.

	Estimate (SE)
—Main effects—	
Intercept	6.756 (0.053)*
List Length	0.439 (0.046)*
Instruction	0.564 (0.036)*
Intact/Reverse	0.138 (0.031)*
Trial	-0.0324 (0.0088)*
Distance	-0.232 (0.014)*
Later-Probe Serial Position (Linear)	-29 (17)*
Later-Probe Serial Position (Quadratic)	-85.5 (7.8)*
—Interactions—	
Instruction × Later-Probe Serial Position (Linear)	-53.7 (4.4)*
Trial × Later-Probe Serial Position (Linear) × Instruction	-6.0 (1.5)*
Intact/Reverse × Later-Probe Serial Position (Linear) × Instruction	-109.5 (7.2)
Distance × Later-Probe Serial Position (Linear) × Instruction	-8.5 (2.1)*
List Length × Intact/Reverse × Instruction × Later-Probe Serial Position (Linear)	95.7 (5.5)*

Table 2.6: The best-fitting LME model for experiment 2 response time. The congruity effect is in bold. The "Estimate" column reports the corresponding regression coefficient, along with its SE (standard error). Significant effects are denoted * - $p < 0.05$. Due to space constraints, this table reports interactions relevant to the Instruction × Later-Probe Serial Position (Linear) only; see supplementary materials Table S1 for the full model.

4 Intact, list length 4 Reverse, list length 8 Intact, and list length 8 Reverse (see Tables S2, S2, S3, S4 and S5 in supplementary materials). The two-way interactions of Instruction × Later-Probe Serial Position (linear component) were significant for all four groups, and the effects were consistent in direction. In addition to the four-way interaction, the congruity effect also interacted with Distance and Trials. The three-way interactions can be understood as increasing Trial number, Distance all selectively facilitating "later" instruction response times at Later-Probe Serial Positions, and having the opposite effect on "earlier" instruction response time at Later-Probe Serial Positions. In other words, the linear Later-Probe Serial Position curve associated with "earlier" instruction is less affected by reverse presentation order, practice effect, and increasing Distance.

Replicating the experiment 1 response time results, the best-fitting LME model also revealed other factors not observable on the data plots, including main effects of List Length, Instruction, Trial and Intact/Reverse. Longer list length, "later" instruction, Reverse presentation order and larger Trial number corresponded with slower response time. The two-way interaction of Instruction × Intact/Reverse was also significant, suggesting a reading-order effect.

In sum, we found a congruity effect on error rate in list length 8, and a response time congruity effect at both list lengths. This challenges the argument that the findings in experiment 1 were a consequence of mixing sub-span lists in with supra-span lists within subjects. Thus, the congruity effect in JORs persists in supra-span lists, despite the differences between experiments 1 and 2, including presentation rate, stimulus materials, and varied versus fixed list lengths.

2.4 Hacker's backward self-terminating search model

The congruity effect may present a new challenge to mathematical models of serial-order memory. Only a few models have been fit to JOR data (e.g., Brown et al., 2000; Hacker, 1980; Lockhart, 1969; McElree & Dosher, 1993). Hacker's (1980) model was designed to explain JOR data with a recency instruction, and makes predictions about both response time and error rate. We ask whether Hacker's (1980) model can already explain the congruity effect in its currently published form. If not, we ask whether the model can be modestly modified to explain the congruity effect.

Hacker (1980) proposed that JOR performance is driven by loss of some items from memory, and backward, self-terminating search of the remaining, available items. The serial-comparison process was assumed to start at the end of the list, progressing toward the beginning (hence, backward), ending when a match to a probe items was found (hence, self-terminating). If an item were "unavailable" due to item loss, the item would not be encountered during search. Probability of a correct JOR (1–Error rate), P_{ij}, can be computed:

$$P_{ij} = \alpha_i + \frac{1}{2}(1-\alpha_i)(1-\alpha_j), \tag{2.1}$$

where i and j are the study–test lags of the more recent and less recent probe items, respectively. α_i is the probability that item i is available in memory, and Hacker treated α_i as free parameters. The first term reflects the case in which the later item is available (a correct response) and the second term represents the case in which both probe items are unavailable, and the response is made by guessing (probability correct=0.5). Hacker went on to model response times on correct trials as follows, assuming that if an item is unavailable, it does not add to the response time.

$$\text{response time}_{ij} = b + \left\{ \alpha_i \left[\left(\sum_{k=1}^{i-1} \alpha_k + 1 \right) s \right] + \frac{1}{2}(1-\alpha_i)(1-\alpha_j) \times \left[\left(\sum_{k=1}^{n} \alpha_k - \alpha_i - \alpha_j \right) s \right] \right\} / P_{ij}, \tag{2.2}$$

where b is a base-level response time for "overhead" processes unrelated to memory and s is the rate to search and compare each available item. The term in the leftmost square bracket represents the expected response time when search ends in a correct match, equal to the summed availability of items less than i that must be compared at rate s ms/item. The sum is incremented by 1 because i must be available to make a correct response (if not a guess). The other term is for the condition when both probes are unavailable, in which case search is exhaustive, summing the availability of all serial positions, excluding the probe serial positions i and j (because they are unavailable), at a rate of s ms/item. The matches and guesses are normalized by the P_{ij} for that comparison.

Note that the same α_i values are used to calculate error rate and response time. For the

Hacker only applied his model to JORs of the last 7 list items. He needed an additional parameter, g, to account for additional searching time towards the beginning of the list after the 7^{th}-back item was reached. Because we applied the model to search through the whole list, we no longer need the "shortcut" parameter g, so we set $g = 0$ to obtain Equation 2.2.

parameter search, we wanted to avoid finding a model that fit "earlier" and "later" instructions individually while failing to capture the difference due to instruction. We therefore opted for a fitness measure that weighted "earlier" data, "later" data and the difference pattern equally. Thus, we fitted Hacker's model by minimizing the summed BIC of "earlier" instruction, "later" instruction and the difference between "earlier" and "later" instruction (both error rate and response time). To compare models from different parameter searches, we recalculated BIC without the redundant "earlier"−"later" terms. We follow the rule of thumb that a change in BIC (ΔBIC) of less than 2 is considered a non-significant difference between models. For error rate, we used the variant of BIC that applies to the special case of least-squares estimation with normally distributed errors on mean performance (D. R. Anderson & Burnham, 2004; Burnham & Anderson, 2002).

Fitting was done in MATLAB (The Mathworks, Inc. Natick, MA) with the simplex algorithm (Nelder & Mead, 1965). With all model fits presented here, the initial parameters were randomly chosen from a range of 0 to 1 for α and 0 to 2000 for b and s and the best-fitting model was the best of 500 executions of the Simplex with different random starting values.

Both list lengths were fit separately. Visual inspection of the simulated data produced by the best-fitting models (Figure 2.7; cf. Figures 2.5 and 2.6) suggests that although the model can reproduce some important features of the data, it does not capture list length 4 error rate pattern well, producing a ceiling error rate for "later" instruction. The model also cannot account for "earlier" instruction response time pattern at both list lengths; in particular, it had trouble producing the primacy-dominant pattern in the response time measure. However, the model produced differences between instructions that resemble the empirical congruity effect qualitatively, and with approximately the same magnitude (cf. Figures 2.3 and 2.5).

In summary, Hacker's backward self-terminating search model ran into problems fitting serial-position effects that have been suggested to reflect forward search, particularly for the list length 4, "earlier" data. Therefore, we next considered whether a forward self-terminating search model would address this limitation.

Note that BIC is a penalized log-likelihood criterion, expressed as $-2(log{-}likelihood) + k * log(n)$, where k represents the number of parameters and n represents the number of observations. Because k and n are constant in our parameter search, the parameter search results should be equivalent to log-likelihood optimization.

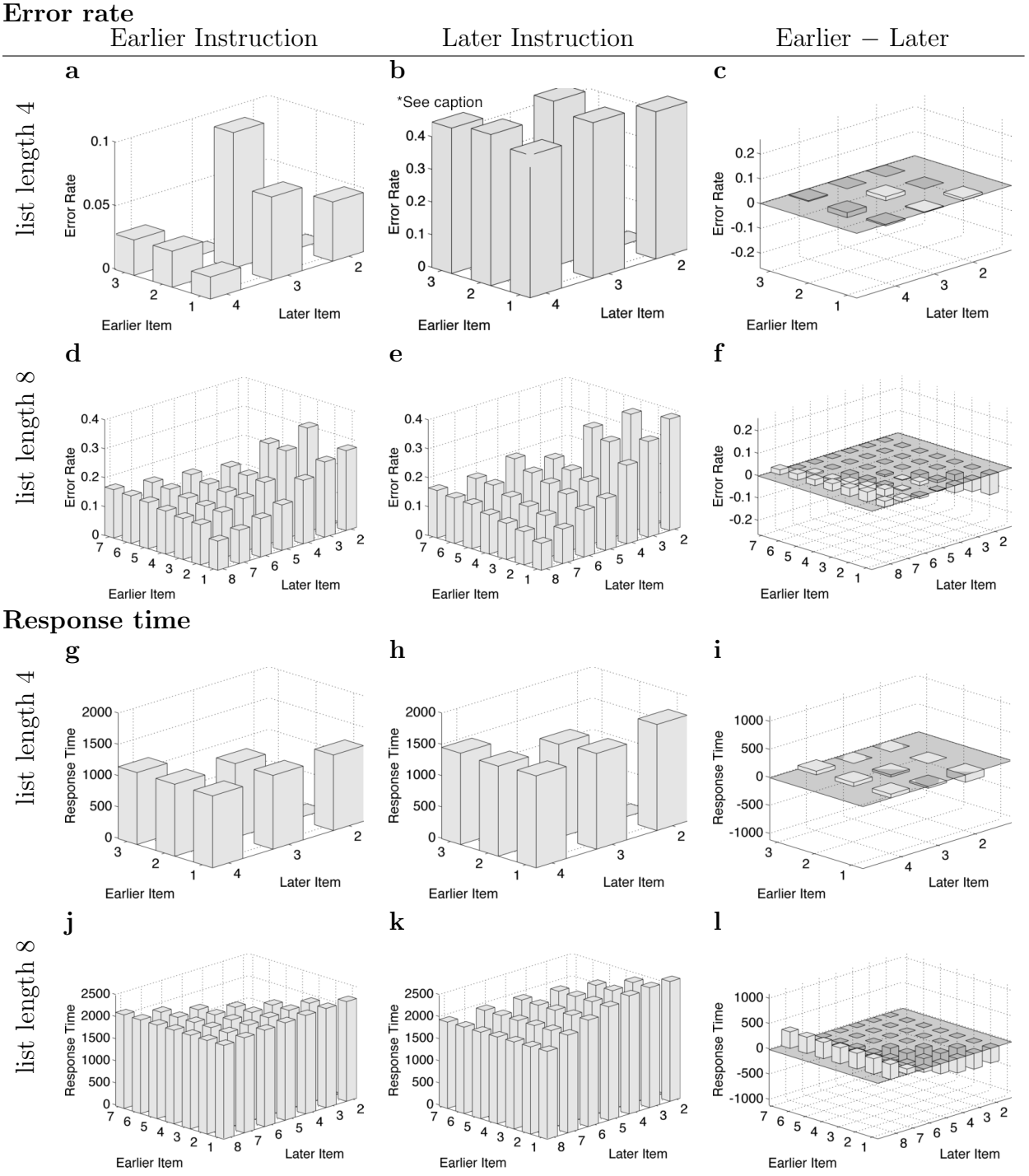


Figure 2.7: Hacker's model error rate (top half) and response time (bottom half), fit to experiment 2, as a function of both probe items' serial position (earlier item and later item, respectively) broken down by list length in rows, and instruction ("earlier", "later" and the difference, "earlier"−"later", corrected for mean response time) in columns. **Note:* The list length 4 error rate "later" instruction is plotted on a different scale than the earlier instruction because this model produced very high values; it could not simultaneously account for both instruction's empirical pattern and their difference pattern.

	Forward		Backward		
list length	b	s	b	s	ΔBIC
list length 4	748.45	316.60	1241.83	0.00	-8.35
list length 8	1882.04	114.96	2069.74	41.01	3.04

Table 2.7: Parameter summary of the Hacker forward versus backward self-terminating search model fitted for "earlier" instruction. Parameters b and s are presented for each model (Forward/Backward) separately (units of ms). Hacker's forward directional search BIC– backward directional search BIC is presented at the last column. Although the best-fitting models were identified using a BIC measure that weighted "earlier," "later" and "earlier"−"later" instructions equally, ΔBIC in this table is computed with "earlier" instruction data only. A negative ΔBIC indicates the forward instruction fit better.

2.4.1 A forward-directed variant of Hacker's self-terminating search model

To implement forward, self-terminating search, for error rate (Equation 2.1) we changed the first α_i to α_j:

$$P_{ij} = \alpha_j + \frac{1}{2}(1 - \alpha_i)(1 - \alpha_j) \tag{2.3}$$

Similarly, for response time (Equation 2.2), we changed the first α_i term to α_j and changed the limits of summation over k. We first asked whether this forward search model would account better for "earlier" instruction data than the backward search model. The best-fitting model parameters from the best-fitting models are summarized in Table 2.7, along with ΔBIC values comparing the forward model and backward models.

The forward model fit "earlier" data better than the backward model for list length 4, but for list length 8, the backward model fit better (lower ΔBIC), and did so by capturing the early-serial-position advantage that presented a problem for the backward model (Figures 2.8). Fitting "earlier" data with the forward model and "later" data with the backward model also improved fits of the congruity effect qualitatively (cf. Figures 2.5 and 2.6).

For more insight, note that for the forward model, "earlier" instruction fit by decreasing α_i over serial position (Figure 2.9a), whereas "later" instruction fit by increasing α_i over serial position (Figure 2.9b). When both "earlier" and "later" instruction fit by the backward model, the α_i values were less steeply sloped for "earlier" than "later" instruction. It may seem surprising that certain values of α_i were near-zero. We understand this as follows. In

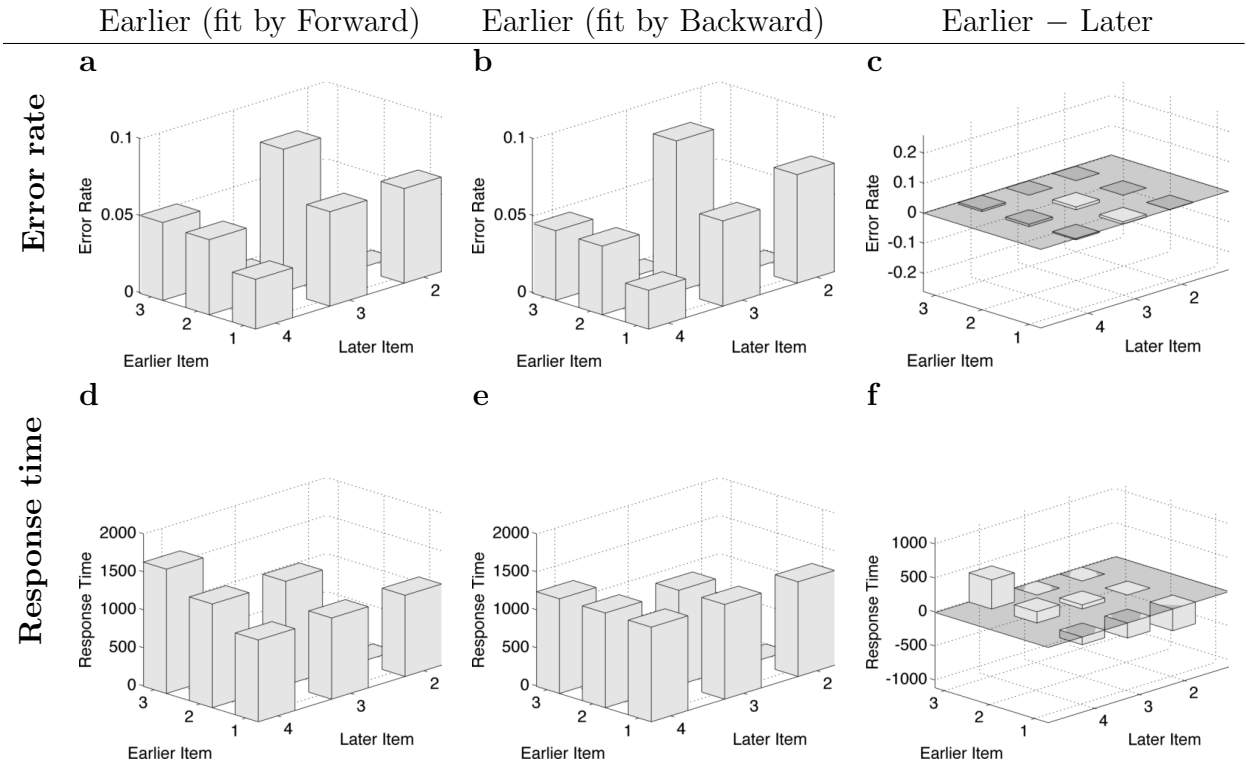


Figure 2.8: The best-fitting hacker's model generated plot using forward direction search for "earlier" instruction (a,d) and backward direction search for "earlier" instruction (b,e). The right-hand column (c,f) represent the hacker's model generated "earlier" − "later" difference pattern when fitting "earlier" instruction with forward directed search and "later" instruction with the backward directional search.

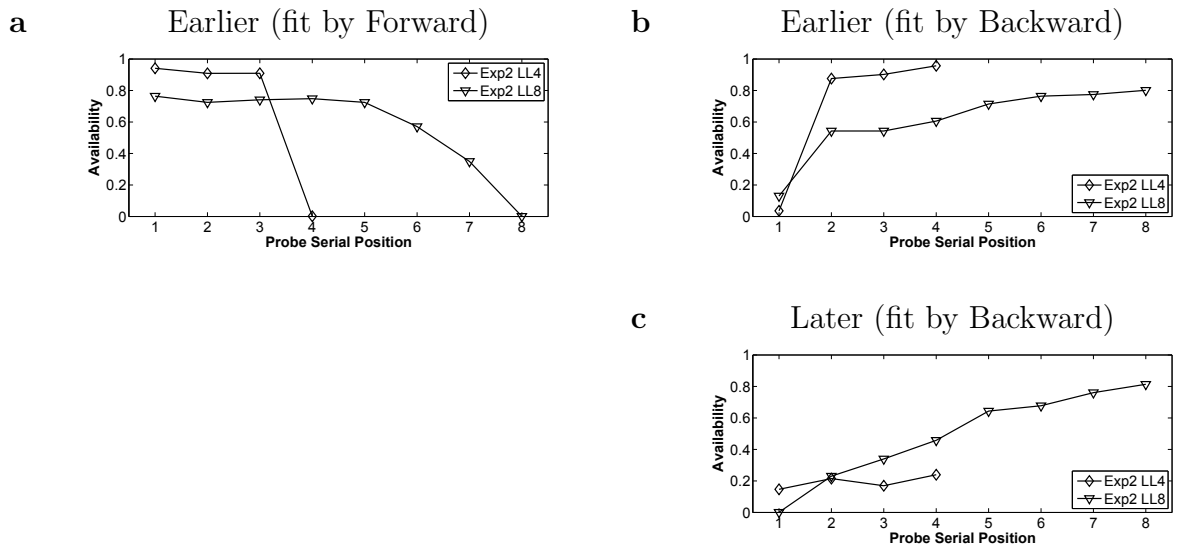


Figure 2.9: Availability (α_i) parameter values plotted as functions of serial position.

"earlier" instruction, the last item of the list can never be a target. Since participants have very good memory of this last item (McElree, 2006), they may easily rule it out as the target and respond correctly. Because Hacker's model selects the item it terminates on as its target, if the $\alpha_{ListLength}$ item were "available," then paradoxically, the response would be incorrect. Thus, it appears that in fitting the model, $\alpha_{ListLength}$ took on a near-zero level as a means of producing very high accuracy for this kind of probe (and likewise for the backward model).

In summary, Hacker's model can fit shorter lists using forward self-terminating search for "earlier" instruction and backward search for "later" instruction. This reversal of search direction does not appear to extend to longer list lengths. For the longer list lengths, direction of search had to be backward for both instructions, but the degrees of freedom contained within the backward, self-terminating search model were sufficient to produce a qualitatively and quantitatively reasonable congruity effect. We discuss alternative model accounts in the General Discussion.

2.5 General discussion

In experiment 1, we found that the congruity effect in the JOR task generalizes to supra-span noun lists, along with the usual distance, primacy and recency effects and an intact/reverse congruity effect. The presence of a congruity effect in error rate suggests that instruction

not only affects order memory retrieval speed, but also the quality of order information that can be retrieved from memory. Experiment 2 replicated the experiment 1 findings, but with consonants and a between-subjects manipulation of list length, suggesting presentation of varied list lengths within subjects does not explain the congruity effect. The fits of Hacker's model and the forward-directed variant suggested that the congruity effect may arise for different reasons at different list lengths; at short list lengths, "earlier" instruction might in fact reverse the direction of self-terminating search, but at longer list lengths, if search is in any sense directional, our model-fits suggest that search is backward for both instructions.

2.5.1 Congruity effect across list length

Our results differ from the list length 4 data reported by Chan et al. (2009) in several ways. Chan et al. (2009) did not find distance effect nor an Intact/Reverse effect, all of which we found in experiment 2, presumably due to higher power and the LME analyses. The finding of long-list-like features like a distance effect may not be surprising, as McElree and Dosher (1993) also found signs of distance effect in relative short lists using a similar JOR response-signal speed-accuracy tradeoff (SAT) procedure. Thus, our findings replicate and extend the congruity effect in sub-span lists reported by Chan et al. (2009).

Extrapolating, one might expect a congruity effect will always be present, even for extremely long list lengths. Alternatively, the congruity effect might become vanishingly small as list length increases. Visual inspection of the data suggests the overall difference in response time remained relatively constant across list lengths. Confirming the visual inspection, LME analysis found the congruity effect did not interact with List Length in both the response-time and error rate data in experiment 1. This suggests that the congruity effect is a general phenomenon that may apply to arbitrarily long lists.

2.5.2 JORs as comparative judgements

Congruity effects similar to ours have been found in closely related paradigms, known as comparative judgements (see Birnbaum & Jou, 1990; Petrusic, 1992; Petrusic, Shaki, & Leth-Steensen, 2008, for reviews), in which a pairwise comparison is made on any of a broad range of stimulus dimensions, including perceptual judgements (e.g., brightness, loudness) and symbolic judgements (e.g., comparing animal size based on animal name). Distance effects, bowed serial position effects and congruity effects were found in our temporal-order

judgement data, and have been commonly found in comparative judgement studies (Banks, 1977). This suggests that JORs may be viewed as a specific instance of comparative judgements, supporting Brown et al.'s (2007) suggestion that temporal order information is processed like magnitude-order information in humans. Thus, congruity effects on JORs may occur for the same reason as they do in other comparative tasks.

Despite the similarities, evidence suggests episodic (temporal order) and semantic judgements of order are not identical. In one study (Jou, 2003), the first nine letters of the English alphabet were the list, and participants were asked to choose either the letter that appears "earlier" or "later" in the alphabet. The 9-item alphabet condition is very similar to our list length 8 JOR task in experiment 2, both with short lists of letters and with "earlier" versus "later" instruction. Jou (2003) found a main effect of instruction, with "earlier" response times faster than "later" response times, but no congruity effect. These results, inconsistent with our findings, could be attributed to the over-learning of the alphabet, or that the forward recall direction is hard to overcome due to the alphabet being highly practised in that direction.

One further reason for caution in relating the memory JOR congruity effect to congruity effects on comparative judgements in that our sub-span results are consistent with sequential, self-terminating search, but to our knowledge, sequential self-terminating search accounts have not been considered for comparative judgements.

2.5.3 Comparison with forward and backward serial recall

The most common procedure used to investigate memory for order is serial recall, where both item and order memory are tested (Kahana, 2012; Murdock, 1974). Could serial recall be the basis of the self-terminating search strategy thought to support JORs? In forward serial recall, participants recall from the beginning toward the end of a list, whereas backward recall starts from the end of the list. At first blush, backward serial recall seems approximately like a mirror-image of forward serial recall, with forward serial recall being dominated by a primacy effect and backward serial recall being dominated by a recency effect (Madigan, 1971; Manning & Pacifici, 1983). Our JOR congruity effect suggests a similar mirroring of serial-position effects as forward versus backward serial recall: "earlier" instruction produced better judgements at earlier serial positions (primacy effect), whereas "later" instruction produced better judgements at later serial positions (recency effect). However, there are

several empirical dissociations that suggest forward and backward serial recall may rely on different cognitive mechanisms (see Richardson, 2007, for a review). Backward serial recall may rely on more visuospatial processing than forward serial recall (Li & Lewandowsky, 1993, 1995; Reynolds, 1997). Thomas et al. (2003) found a response time pattern that suggested simple sequential search of the items in forward recall but for backward recall, a U-shaped response time curve suggested participants may have used multiple forward recalls when recalling backward.

Another interesting set of findings that may inform our results comes from a comparison of free recall with forward serial recall (Ward, Tan, & Grenfell-Essam, 2010). Because free recall does not dictate order of report, participants are free to initiate recall at any serial position. Ward et al. (2010) found that for shorter list lengths, the free-recall order resembled their forward serial-recall results; thus, participants prefer to recall short lists in the forward direction. In contrast, at long list lengths, participants chose to initiate recall with one of the last four items, which, although not identical, is more like backward than forward serial recall. This may indicate that a forward search strategy is available and convenient for JORs, but more so for short than long lists, which is consistent with our model fits. Thus, JORs might be carried out using a covert serial-recall-like strategy, especially at shorter list lengths. This hypothesis leads to interesting, testable predictions. If JORs rely on serial recall, then the manipulations that previously dissociated forward from backward serial recall (Beaman, 2002; Reynolds, 1997; Li & Lewandowsky, 1993, 1995; Madigan, 1971; Manning & Pacifici, 1983; Thomas et al., 2003) should produce analogueous dissociative effects on JOR behaviour comparing "earlier" versus "later" instructions.

2.5.4 Models of order-memory and the congruity effect

Although a full consideration of the implications of our findings for models of order-memory is beyond the scope of this paper, there are some points we can make clearly that speak to the inadequacies of current models and possible future directions for model development in light of our findings.

We first consider Hacker's (1980) model, an implementation of sequential, self-terminating search. We considered this model in depth because it has been successfully applied, several times, to JOR data. We asked if this pre-existing model could already produce a congruity effect. Although it could not, an adaption of Hacker's model could capture the congruity

effect in sub-span lists— namely, assuming forward directional search for "earlier" instruction and backward directional search for "later" instruction. For short lists, then, there may be no effect of instruction on the underlying processes generating the behaviour, apart from a reversal of search-direction. However, the forward directional search model was not compatible with "earlier" instruction data of the supra-span lists, even despite this model's large number of degrees of freedom, which becomes larger as list length increases. This may indicate that a single explanation of the congruity effect is not possible for both short and long lists. Rather, it may be that the mechanism shifts at some critical list length— but if so, it remains to be determined what principle governs that switch in search direction. Finally, it is important to note that, because we only fit a single model to our data, that does not mean that the model is confirmed. It is quite plausible that a different model (possibly variants of the models we review in this section) would produce a better fit, both quantitatively and qualitatively. The level of success of this model, therefore, should not be taken as support for this particular model over other models.

At first glance, a self-terminating search mechanism presented in Hacker's (1980) model could be compatible with other models of order memory applied to serial recall. For example, an associative chaining model, where each item is associated with the previous item in the list to form a chain (e.g., Kleinfeld, 1986; Lewandowsky & Murdock, 1989; Riedel, Kühn, & van Hemmen, 1988; Sompolinsky & Kanter, 1986; Wicklgren, 1966), and positional coding models, where item position is used to probe each item (e.g., Burgess & Hitch, 1999; Henson, 1998). Both chaining and positional coding mechanism could be used to model self-terminating search. However, a key assumption of Hacker's model differs from chaining and positional coding models: that an item can be skipped without any impact on response time, which is how Hacker's model produces a distance effect. To our knowledge, both chaining and positional coding models have not been implemented in such a way that they save processing time for a missed item. Chaining models may handle a missed item by probing with the previously retrieved vector even if the correct response could not be made (e.g., Lewandowsky & Murdock, 1989). Positional coding models continue to probe with the subsequent position, regardless of accuracy of the previous recall (e.g., Burgess & Hitch, 1999; Henson, 1998). Thus, current models of serial-order memory would need to be modified to incorporate Hacker's mechanism.

Even if an account based on Hacker's model is correct, this model was only developed

to explain the JOR task; in its current formulation, it does not do other order-memory tasks, like serial recall. Rather than start with a model of JORs and figure out how to develop it into a full-fledged memory model, one could consider models that were designed to explain serial-recall data, and ask how such models might handle the JOR task. OSCillator-based Associative Recall (OSCAR; Brown et al., 2000) is a model of serial recall that has actually been fit to JOR data with some success. In this model, items are assumed to be associated with the state of an internal context signal (activation values of a bank of sine-wave oscillators), and retrieval of items requires re-instatement of the context. The authors applied OSCAR to the JOR task (Hacker's 1980 data) by probing with the end-of-list context vector. More recent items tend to be more similar to the end-of-list context. The strongest activated list item was compared to the probe items; if a match was found, the search terminated; if no match was found, the next-highest activated item was considered next, and so on. It is not obvious to us how the congruity effect could be explained with this approach. At the very least, to explain the sub-span "earlier" data, the model might need to be able to substitute the start-of-list context, and the congruity effect in supra-span lists, dominated by an overall recency effect, would still remain to be explained.

TODAM is another model that has been fit to JOR data (Murdock, Smith, & Bai, 2001). In this version of the model (TODAM2), recency was judged based on strength of the item-memory terms (not the association terms that are used in serial-recall), and more recent items had greater strength. This could explain serial-position effects that are dominated by recency, such as we found in supra-span lists, but it is not obvious how this mechanism could be adapted to produce the primacy-dominant pattern found for list length 4. Furthermore, the congruity effect in supra-span lists would still need to be explained. Finally, TODAM was only implemented for error rates and not response times, so additional modifications would be necessary to explain the response-time data.

SIMPLE, a scale-invariant model that assumes that memory is driven by discriminability of presentation times of items (Brown et al., 2007), produces bow-shaped serial-position effects and a distance effect, but it remains unclear how the model might account for the congruity effect. One might assume different instructions can systematically distort the representation of time either directly, or influencing judgements on a separate, serial-position dimension. An interesting possibility is that the congruity effect might be produced by participants encoding list position differently, depending on instruction (Neath & Crowder,

1996); for example, with the first item first for "earlier" instruction, and the last item first for "later" instruction. Although promising, the current version of SIMPLE does not model response time data, which means more work is required to adapt SIMPLE to explain the full pattern of JOR data reported here.

In short, to our knowledge, no model of serial recall in its current form is sufficient to explain the JOR congruity effect across list lengths.

2.6 Conclusion

In sum, the pattern of both speed and errors depends on how the order-judgement question is asked. If the target is the earlier item, judgements are better at earlier serial positions, whereas if the target is the later item, judgements are better at later serial positions, reminiscent of congruity effects found in comparative judgements. A self-terminating search model could account for sub-span data by a reversal of search direction between instructions, but longer-list data demanded a different account (both backward-search). Direct-access accounts hold promise, but it is unclear how they could capture the full pattern of serial position effects on both error rate and response time measures, across list lengths. Thus, although instruction has a similar effect across list length, either the underlying mechanisms driving the congruity effect change with list length, or a unified account may need to combine elements of both types of model.

2.7 Additional analysis

Here we present additional modelling analysis, fitting the JOR data in Chapter 2 to SIMPLE, and further compare SIMPLE with Hacker's (1980) model. This section is not part of the published version of the Liu et al. (2014) study.

2.7.1 SIMPLE

SIMPLE is a scale-invariant model that assumes that memory is driven by discriminability of presentation times of items (Brown et al., 2007). Time is assumed to be perceived logarithmically relative to time of test. This produces a recency effect, because more recently presented items are more discriminable from one another, being relatively less compressed by the log-transform. Also, when the serial positions of probe items are farther apart from each other, the memories are more discriminable, explaining the distance effect. SIMPLE also produces both a primacy and an additional recency effect due to edge effects: items at the beginning or end of a list have no competing items earlier or later, respectively; this means primacy and recency items are more discriminable than items in the middle of the list.

We first asked if a straight-forward application of the existing formulation of SIMPLE might already account for the difference we observed due to instruction. We considered the possibility that SIMPLE could adjust its main free parameter, c (see below), to fit each instruction data set on its own. This difference in fit might already resemble our congruity effect. The values along the psychological dimension (in our case, item presentation time), are denoted M_i and M_j for a given pair of items, with time of presentation i and j. Similarity, $\eta_{i,j}$, between the two item times is:

$$\eta_{i,j} = e^{-c|M_i - M_j|}, \tag{2.4}$$

where c controls how mutually confusible presentation times are. Although motivated by research on absolute and relative identification, To our knowledge, SIMPLE has not yet been applied to the JOR task (i.e., relative identification based on remembered time), but it was suggested (Brown et al., 2007). We had to incorporate additional assumptions about how similarity values are used to generate the JOR response. We assume the participant attempts to retrieve the absolute position of each probe item and compares the two; for

the "later" instruction, the model chooses the item retrieved with later position and for the "earlier" instruction, the model chooses the item retrieved with the earlier position. Because of temporal confusibility, there is a probability that the probe items' positions will be transposed, producing an error response (Hacker, 1980; Lockhart, 1969). If the retrieved positions, by chance, are equal, we assume the model guesses with 50% accuracy. First, we compute for each probe-item's serial position, i, the probability that the retrieved position is r, where r ranges across all list positions:

$$P(r|i) = \frac{\eta_{i,r}}{\sum_{k=1}^{n} \eta_{i,k}} \tag{2.5}$$

where n is the LL. Note that similarity between positions is what determines errors in retrieved position. When the distribution of retrieved positions is calculated for both probes, we compute ER by calculating, for a given probe (i, j), the probability that i and j are retrieved in the correct order. Note that even if both probes are retrieved in erroneous positions, there is a large probability that their *relative* positions will be left intact.

$$P(\text{respond "i < j"}|i,j) = \sum_{a=1}^{n-1} \left[P(a|i)) \sum_{b=n-a}^{n} P(b|j) + 0.5 \sum_{c=1}^{n} P(c|i)P(c|j) \right] \tag{2.6}$$

where $P(a|i)$, $P(b|j)$, $P(c|i)$ and $P(c|j)$ are obtained with Equation 2.5. The last term within the brackets is the 50% guess rate when the two retrieved positions happen to be the same. Error rate, ER, as a function of the probe, (i, j), is:

$$ER(i,j) = \begin{cases} P(\text{respond "i < j"}|i,j) & i > j \\ 1 - P(\text{respond "i < j"}|i,j) & i < j \end{cases} \tag{2.7}$$

Visual inspection of the data (Figures 2.5) suggests that there is a base level of ER that applies to all probes, even the easiest. This is what one would expect if participants were guessing on some proportion of trials. We therefore added a guessing mechanism; parameter g denotes the proportion of guess trials. The adjusted error rate, ER', is:

$$ER'(i,j) = ER(i,j)(1-g) + 0.5g, \tag{2.8}$$

where $ER(i, j)$ is obtained with Equation 2.7.

The best-fitting models with c and g as the two free parameters are plotted for each instruction (Figure 2.10), and the corresponding parameter values are summarized in Table 2.8. This version of SIMPLE, which we call $c + g$, accounted for the key features of the

	Earlier		Later	
LL	c	g	c	g
LL4	14.81	0.09	19.14	0.11
LL8	7.21	0.27	5.01	0.22

Table 2.8: Parameter summary for the $c + g$ SIMPLE model fitted for the "earlier" and the "later" instruction and "earlier" – "later" difference simultaneously.

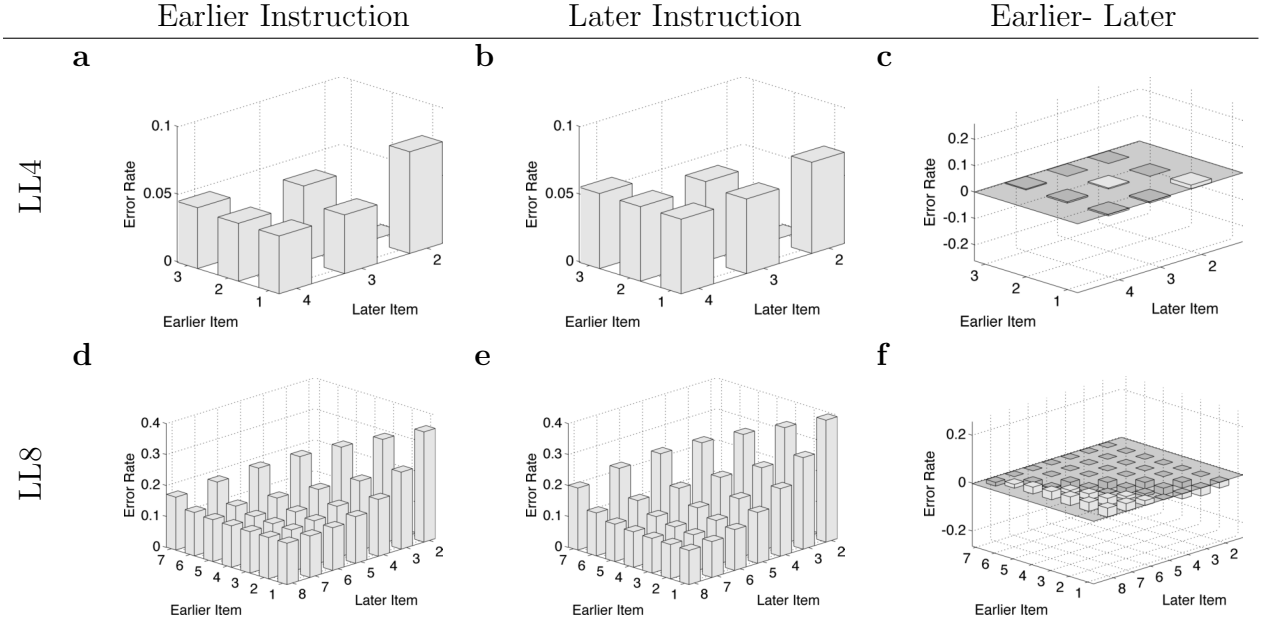


Figure 2.10: SIMPLE $(c + g)$ model fits of ER (Experiment 2) as a function of both probe items' SP (earlier item and later item, respectively) broken down by LL in rows, and instruction ("earlier", "later" and the difference, "earlier"–"later", corrected for mean RT) in columns.

data: primacy, recency, and distance effects, as claimed by Brown et al. (2007). However, qualitatively, the difference pattern did not resemble the empirical data and the magnitude was much smaller (Figure 2.10c,f).

2.7.2 SIMPLE + Temporal gradient model

To enable SIMPLE to better account for the congruity effect, we added a gradient to the temporal code, t, that could make earlier list items more discriminable for the "earlier" instruction and later list items more discriminable for the "later" instruction. The gradient represents a psychological distortion of the temporal dimension, rather than combining two

dimensions as suggested by Brown et al. (2007). We have attempted to fit our data to a two-dimensional version of SIMPLE, by adding an additional dimension to represent a directional bias. However, we could not obtain a good fit of our data, as the added dimension is either weighted 100% or 0% by the best fitting model. The two dimensional model may need further development to account the JOR data set.

The added gradient is thus a list of increasing numbers with range from –1 to 1 and mean of 0 (e.g., [–1 –0.5 0 0.5 1]) - for the example of a 5-item list:

$$t_i - t_{i+1} = 1 + \tau \frac{i - (n+1)/2}{n - (n+1)/2} \tag{2.9}$$

where t_i is the recency of item i. $i-(n+1)/2$ represents the contribution of the gradient; its steepness is controlled by τ.

The best-fitting SIMPLE + temporal gradient models with c, g and τ as free parameters were plotted for each instruction in Figure 2.11 and the corresponding parameter values were summarized in Table 2.9. τ was more negative in the "later" instruction (indicating greater discriminability toward the start of the list) than the "earlier" instruction in LL8; however, τ was more positive in LL4.

With the additional gradient, the congruity difference pattern was captured very well qualitatively, with the magnitude of the difference-plot values being larger, and closer to the values of the data, than the $c + g$ model. ΔBIC relative to the $c + g$ model revealed a significantly better fit for LL8, but worse for LL4.

In summary, a model close to published forms of SIMPLE ($c + g$) could not account for the magnitude of the congruity effect, but could with an additional gradient parameterized by τ at LL8. τ did not help the model fit LL4 better.

2.7.3 Comparing Hacker's model with SIMPLE

After having evaluated each model individually, we now compare our best versions of each model to determine whether one model provides a better account of the full pattern of the data. Since SIMPLE does not model RTs, we can only compare the models based on their fits to the ER data alone (this entailed refitting the Hacker-based models to the ER data without RT as a constraint). BIC and ΔBIC are summarized in Table 2.10. Hacker's model

	Earlier			Later			
LL	c	g	τ	c	g	τ	ΔBIC
LL4	11.42	0.08	0.27	12.67	0.10	0.40	-3.32
LL8	6.01	0.26	0.38	6.01	0.24	-0.20	6.17

Table 2.9: Parameter summary of the temporal gradient model. The ΔBIC column is the BIC for SIMPLE $(c+g)$ minus the BIC for SIMPLE + temporal gradient. A positive ΔBIC would indicate a better fit of the temporal-gradient than the $(c+g)$ model.

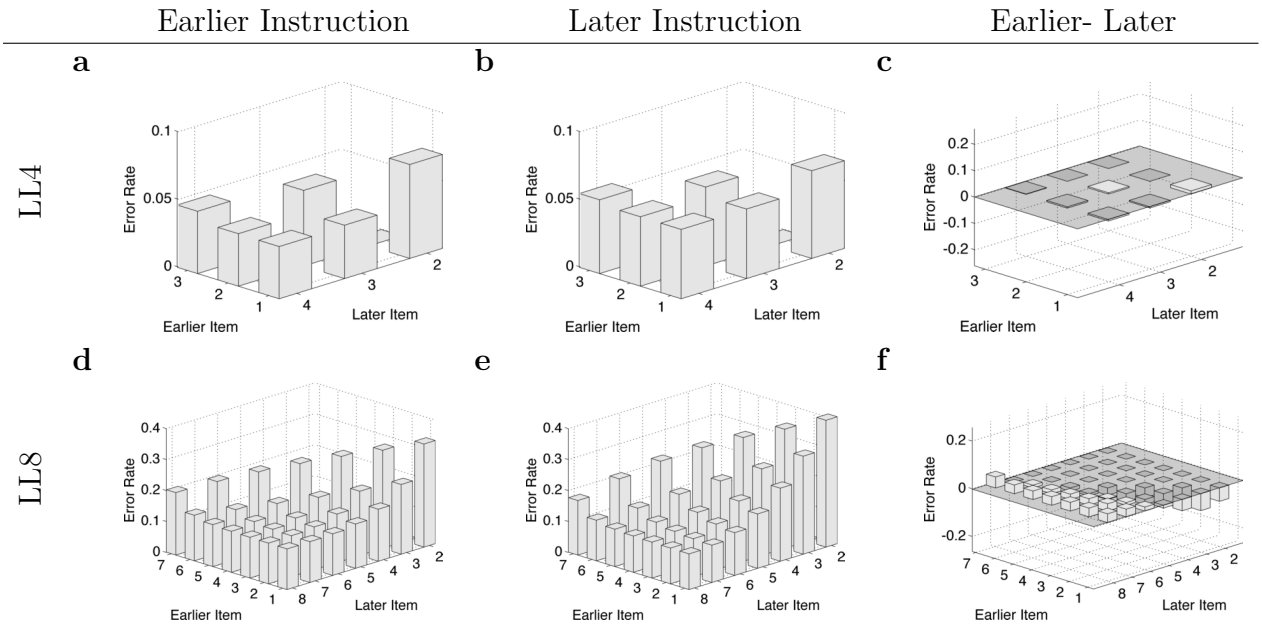


Figure 2.11: SIMPLE + Temporal gradient model fit ER (Experiment 2) as a function of both probe items' SP (earlier item and later item, respectively) broken down by LL in rows, and instruction ("earlier", "later") and the difference, "earlier"−"later", corrected for mean RT) in columns.

	Hacker's Original	Forward Early	SIMPLE with temporal gradient
LL4	-493.91	**-509.27**	-142.34
LL8	-484.94	-445.95	**-534.91**

Table 2.10: Summary of models' BIC values. "Hacker's original" is Hacker's model fitting backward directional search for both the "earlier" and "later" instructions. "Forward Early" is Hacker's model fitting forward directional search for the "earlier" instruction and backward directional search for the "later" instruction. "SIMPLE with temporal gradient" is the SIMPLE model with addition of temporal gradient parameterized by τ. Boldface indicates the lowest BIC for the LL.

using forward search for the "earlier" instruction was best at fitting LL4. For LL8, SIMPLE with temporal gradient clearly outperformed the Hacker models. These results suggest that whereas a directional self-terminating search model can fit shorter lists well, SIMPLE with a temporal gradient model can account for the longer lists better.

One important caveat should be noted: as LLs increases, the number of free parameters α_i increases, which will increase BIC, perhaps accounting for some of the poorer performance of the Hacker model variants at longer LLs. Also, the fits of Hacker's model suggest a switch in strategy between forward and backward directional search somewhere along the LL continuum, so it is less parsimonious for this reason, without a principle governing at which LLs the directional switch is required.

In summary, both models have limits in explaining the JOR congruity effect, with each model performing better at a different LL. This suggests that a different mechanisms produces a similar congruity effect for short versus long lists, or that if a single mechanism can explain congruity effects across list lengths, it remains to be identified.

2.8 Supplementary Materials

Following the convention of the lmer function output format, we report the best-fitting LME summary tables for experiment 2 response time in Table S1, and separate fits for each list length and Intact/Reverse combinations for experiment 2 response time (Table S2, S3, S4, S5). Due to table width constraint, we use abbreviations in this section: Intact/Reverse (IR), List length (LL), and Later-Probe Serial Position (LPSP). Note that "earlier" and "later" instruction were presented separately in summary tables if the factor interacting with Instruction has no main effect.

The Instruction × quadratic component of Later-Probe Serial Position × List Length three-way interaction from the best fitting LME model of experiment 2 response time is presented in Figure S1. The Instruction × linear component of Later-Probe Serial Position × Distance three-way interaction from the best fitting LME model of experiment 2 response time is presented in Figure S2.

		Estimate (SE)
Main effects		
	Intercept	6.756 (0.053)*
	ListLength	0.439 (0.046)*
	Trial	-0.032 (0.009)*
	IR	0.138 (0.031)*
	Distance	-0.232 (0.015)*
	LPSP(Linear)	-29.38 (16.88)*
	Instruction	0.564 (0.036)*
	LPSP(Quadratic)	85.48 (78.16)*
Interactions		
	LL × IR	-0.024 (0.026)
	LL × Distance	0.022 (0.008)*
	Trial × IR	-0.026 (0.006)
	Trial × Distance	-0.04 (0.004)
	LL × LPSP(Linear)	59.92 (13.73)
	IR × Distance	0.036(0.008)
	LL × Instruction	-0.059(0.038)
	Trial × LPSP(Linear)	1.961(1.011)
	LL × LPSP(Quadratic)	39.97(6.263)*
	IR × LPSP(Linear)	-0.512(11.11)
	Trial × Instruction	-0.091(0.013)*
	Distance × LPSP(Linear)	58.22(3.866)*
	Trial × LPSP(Quadratic)	-1.755(0.653)*
	IR × Instruction	-0.566(0.020)*

IR × LPSP(Quadratic)	13.30(4.987)
Distance × LPSP(Quadratic)	-0.349(0.936)
Distance × Instruction	0.208(0.008)
LPSP(Linear) × Instruction	-53.72(4.383)*
Instruction × LPSP(Quadratic)	44.14(2.922)*
LL × IR × LPSP(Linear)	-8.331(9.025)
LL × Distance × LPSP(Linear)	-33.53(3.090)*
LL × IR × Instruction	0.297(0.019)*
LL × IR × LPSP(Quadratic)	0.064(4.011)
LL × Distance × Instruction	-0.080(0.010)*
Trial × IR × Instruction	0.040(0.008)*
IR × Distance × LPSP(Linear)	-4.664(1.224) *
Trial × Distance × Instruction	0.000(0.006)
Trial × LPSP(Linear) × Instruction	-6.025(1.447)*
IR × LPSP(Linear) × Instruction	-109.5(7.169)*
Trial × Instruction × LPSP(Quadratic)	4.668(0.941)*
Distance × LPSP(Linear) × Instruction	-8.512(2.085)*
IR × Instruction × LPSP(Quadratic)	-63.36(3.404)*
LL × IR × LPSP(Linear) × Instruction	95.67(5.525)*
LL × IR × Instruction × LPSP(Quadratic)	30.66(2.431)*

Table S1: The best-fitting LME model for experiment 2 response time. The congruity effect is in bold. The "Estimate" column reports the corresponding regression coefficient, along with its SE (standard error). Significant effects are denoted * - $p < 0.05$.

		Estimate (SE)
Main effects		
	Intercept	7.437(0.037)*
	Trial	-0.0581(0.009)*
	Instruction	0.233(0.052)*
	LPSP(Linear)	20.99(2.097)*
	LPSP(Quadratic)	-15.13(1.583)*
Interactions		
	Trial × Instruction(Later)	-0.060(0.013)*
	Instruction(Earlier) × Distance	-0.142(0.012)*
	Instruction(Later) × Distance	-0.088(0.012)*
	LPSP × Distance	8.322(1.647)*
	Instruction × LPSP(Linear)	-31.82(2.772)*

Table S2: The best-fitting LME model for experiment 2 list length 8 response time with intact presentation order. The congruity effect is in bold. The "Estimate" column reports the corresponding regression coefficient, along with its SE (standard error). Significant effects are denoted * - $p < 0.05$.

		Estimate (SE)
Main effects		
	Intercept	7.538(0.036)*
	Trial	-0.085(0.007)*
	Instruction	0.076(0.051)
Interactions		
	Instruction(Earlier) × Distance	-0.080(0.006)*
	Instruction(Later) × Distance	-0.028(0.006)*
	Instruction(Earlier) × LPSP(Linear)	13.22(2.129)*
	Instruction(Later) × LPSP(Linear)	-21.38(2.147)*
	Instruction(Earlier) × LPSP(Quadratic)	-5.589(1.649)*
	Instruction(Later) × LPSP(Quadratic)	-17.47(1.649)*

Table S3: The best-fitting LME model for experiment 2 list length 8 response time with reverse presentation order. The congruity effect is in bold. The "Estimate" column reports the corresponding regression coefficient, along with its SE (standard error). Significant effects are denoted * - $p < 0.05$.

		Estimate (SE)
Main effects		
	Intercept	6.429(0.049))*
	Trial	-0.044(0.007)*
	Distance	-0.177(0.016)*
	LPSP(Linear)	-102.9(13.18)*
	Instruction	0.025(0.048)
	LPSP(Quadratic)	-133.0(6.713)*
Interactions		
	Trial × LPSP(Linear)	6.141(2.212)
	Trial × Instruction	-0.094(0.010)*
	Distance × LPSP(Linear)	115.7(8.609)*
	Distance × Instruction	0.244(0.021)*
	LPSP(Linear) × Instruction	-183.0(8.594)*
	Trial × LPSP(Linear) × Instruction	-20.81(3.254)*
	Distance × LPSP(Linear) × Instruction	-54.14(10.59)*

Table S4: The best-fitting LME model for experiment 2 list length 4 response time with intact presentation order. The congruity effect is in bold. The "Estimate" column reports the corresponding regression coefficient, along with its SE (standard error). Significant effects are denoted * - $p < 0.05$.

		Estimate (SE)
Main effects		
	Intercept	6.922(0.032)*
	Trial	-0.072(0.007)*
	Distance	-0.188(0.014)*
	Instruction	-1.170(0.066)
Interactions		
	Trial × Distance	-0.043(0.009)
	Trial × LPSP(Linear)	12.56(2.648)*
	Trial × Instruction	-0.047(0.011)
	Distance × LPSP(Linear)	54.87(7.021)*
	Distance × Instruction	0.241(0.015)*
	Instruction(Later) × LPSP(Linear)	-480.2(17.91)*
	Instruction(Earlier) × LPSP(Quadratic)	-55.92(2.630)*
	Instruction × LPSP(Quadratic)	-217.4(8.727)*
	Trial × Distance × Instruction	0.062(0.013)
	Trial × Instruction × LPSP(Linear)	-27.39(3.805)*

Table S5: The best-fitting LME model for experiment 2 list length 4 response time with reverse presentation order. The congruity effect is in bold. The "Estimate" column reports the corresponding regression coefficient, along with its SE (standard error). Significant effects are denoted * - $p < 0.05$.

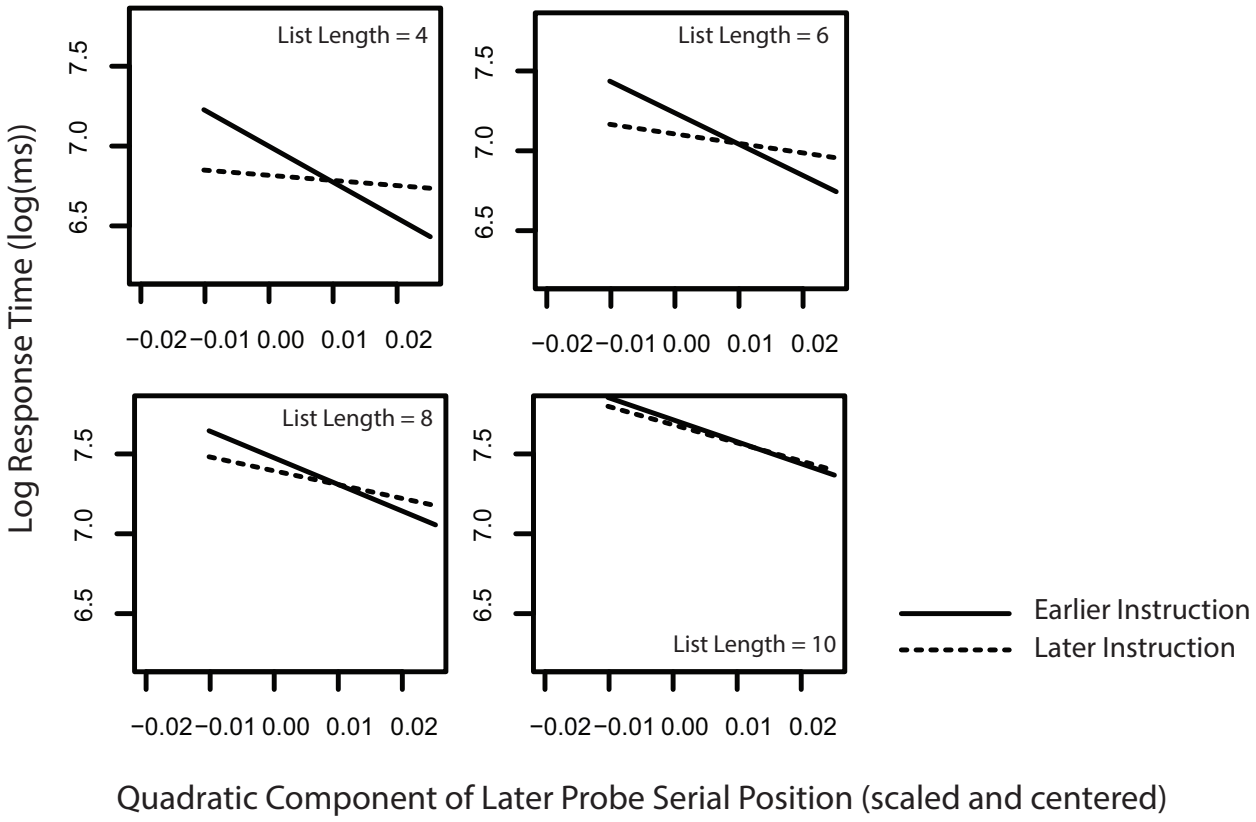


Figure S1: Best fitting LME plot of Instruction × quadratic component of Later-Probe Serial Position × List Length interaction. Instruction × quadratic component of Later-Probe Serial Position is plotted at all levels of List Length.

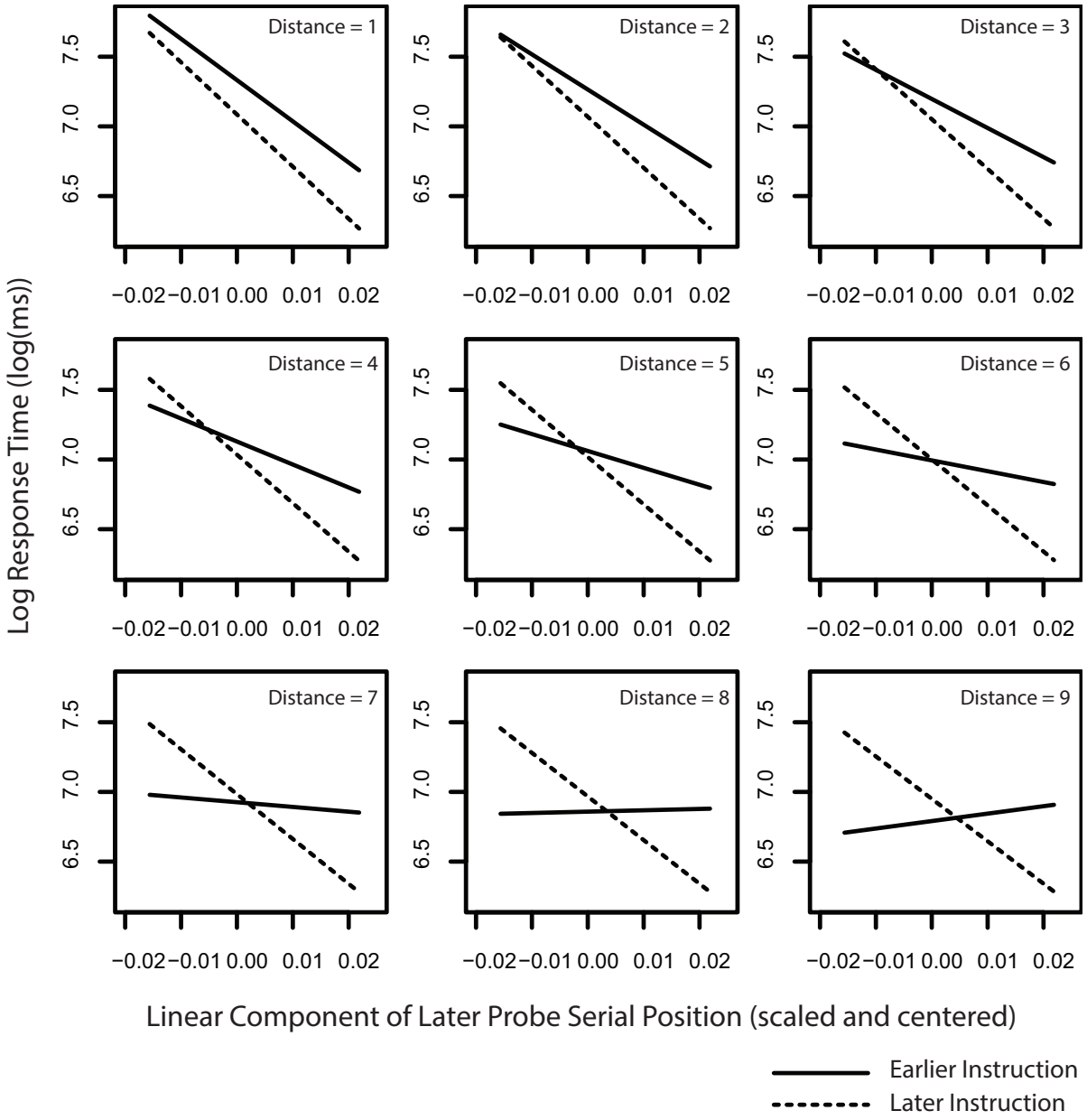


Figure S2: Best fitting LME plot of the interaction of Instruction × linear component of Later-Probe Serial Position × Distance. Instruction × linear component of Later-Probe Serial Position is plotted at all levels of Distance.

Chapter 3

Congruity effect in alphabetical order judgements

3.1 Introduction

Serial-order memory is critical for a very broad range of human behaviour (e.g., Lashley, 1951; Nipher, 1878). One way to test order-memory is to ask participants to judge the relative order of two items from a sequence. For example, after studying a list ABCD, one could be asked "which item is more recent? (B or D)" This kind of two-alternative forced choice, relative temporal-order judgement has been called a judgement of relative recency (Hacker, 1980; Muter, 1979; Yntema & Trask, 1963). We use the more generic term, judgements of relative order (JOR) to include order judgements based on instructions that do not refer to recency (Chan et al., 2009; Liu et al., 2014). Memory researchers and modellers interested in the JOR procedures have made little contact with a closely related paradigm, comparative judgements. Comparative judgement research typically examines comparisons of perceptual judgements of physical magnitudes, such as judgements of differences between luminance levels (Cattell, 1902), pitch (Audley & Wallis, 1964; Banks & Root, 1979), size and weight (Masin, 1995; Paivio, 1975). This approach was later extended to the symbolic domain, including judgements of size, such as the concept of an elephant versus a mouse (e.g., Banks et al., 1983; Cech & Shoben, 2001), and subjective dimensions like preferences (Birnbaum & Jou, 1990), relative age (Ellis, 1972), probability of events (Marks, 1972) and demographic knowledge (Schweickart & Brown, 2013). In this chapter, we consider the possibility that memory JOR behavior is best understood as an instance of the comparative judgement task (Brown et al., 2007).

Three key properties are found in comparative-judgement data (see Petrusic, 1992; Leth-Steensen & Marley, 2000, for reviews): a) a distance effect, characterized by a decrease in response time when the distance between the probe items increases; b) an inverted U-shaped serial position response time or error rate serial position curve, with poor performance at middle of the list and enhanced performance at either end of the list; and c) a congruity effect, characterized by a decrease in response time and error rate when the wording of the question is congruent with the probe on a relevant dimension. Both a distance effect and an inverted U-shaped serial position effect have been found in judgements of relative recency (Hacker, 1980; Muter, 1979; Yntema & Trask, 1963). Chan et al. (2009) found a congruity effect with short lists of consonants (list length = 3, 4, 5, 6), where asking "which item came earlier" selectively enhanced relative-order judgement speed toward the beginning of the list, and asking "which item came later" selectively enhanced judgements toward the end of the list. In chapter 2 (Liu et al., 2014) we showed that the congruity effect generalized to longer temporally presented lists (8 consonants and 4, 6, 8 and 10 nouns) and could be seen in error-rate, as well as response-time data.

To test whether the congruity effect is a general phenomenon of memory judgements of order, spanning from semantic to temporal (episodic) lists, we examined the English alphabet as a test case. The alphabet is a well practiced, long list, with very good item and order encoding (Klahr, Chase, & Lovelace, 1983). If we find a congruity effect, it would help clarify the previous findings on alphabetical judgements, and would suggest a continuum from judgements of relative temporal order to judgements of semantic-memory order. Moreover, it would support attempts to explain memory judgements together with comparative judgements beyond the domain of memory (e.g., Brown et al., 2007).

Despite the common features of JOR and comparative-judgement data, there are some differences to consider. First, comparative judgements typically start by training participants to a ceiling accuracy criterion, or test knowledge mastered prior to the experiment (e.g., Birnbaum & Jou, 1990; Jou, 2005). This makes the memory more likely to be semantic-like than episodic-like, and more importantly, ceiling performance limits the measure to response time only. Error-rate congruity effects are thus rare in comparative judgements (but see Petrusic, 1992). We wondered whether the same error-rate congruity effect could be found for semantic-memory ordered lists, as it has been for temporally presented lists. Although the alphabet is a list mastered pre-experimentally, we thought it might be long

enough that with a large enough sample size, a congruity effect might be detectable in error-rate data. The finding of error rate congruity effect would help us to better understand this novel phenomena. We suggest the lack of error rate congruity effect in previous comparative judgements literature could be caused by non-linear speed-accuracy tradeoffs, where speed trade off with accuracy differently across serial positions.

There are some published studies of order-judgements of the alphabet, but the results appear inconsistent with the finding of a congruity effect in temporal JORs (Chan et al., 2009; Liu et al., 2014). Analyzing relative-order judgements involving the first 9 letters of the English alphabet, Jou (2003) found a main effect of instruction on response time, but no congruity effect (i.e., no interaction between serial-position and instruction). Using more of the alphabet (excluding the first and last three letters), Jou and Aldridge (1999) also failed to report a congruity effect. However, the congruity effect has been reported in judgements among letter-triplet probes extracted from the whole alphabet (Jou, 1997). The lack of congruity effect in two-alternative forced choice alphabet judgement may have been due to insufficient power, or testing only a limited portion of the alphabet. We further suggest the inconsistency of findings may have been due to task differences between two-alternative forced choice task and multiple-alternative forced choice tasks (Jou, 1997). For the two-alternative forced choice task, asking which letter came earlier is a logical equivalent of asking which letter came later, as if probe A is earlier, probe B must be later, and vice versa. Thus, it is possible that participants internalize this relationship and utilize it for the JOR strategy. Unlike the two-alternative forced choice task, instruction such as "chose the earliest/latest letter" in multiple-alternative forced choice tasks are not logical equivalents. For example, for a three-alternative forced choice task, if the "earlier" strategy is used for the "later" judgement, participants need to identify the earliest item of the list, then determine which item is earlier for the remaining two items in order to judge which item is the latest. This process demands at least twice the cognitive load as switching to a different strategy, and the cognitive process is more elaborate as the number of choices increases, thus creating a incentive for participants to switch strategy. For two-alternative forced choice task, participants may switch strategies voluntarily, without any efficiency incentive. This hypothesis would be supported if we fail to find a congruity effect using a two-alternative forced choice task with a large sample size and a full list of the alphabet.

Many findings would lead one to expect a congruity effect in JORs of the English alphabet.

Directly relevant to semantic memory, a congruity effect has been found with response-time measures in shorter, highly practiced lists, such as months of the year (Gelinas & Desrochers, 1993). A response-time congruity effect has also been demonstrated for order-judgement of the events in a story script (Wyer et al., 1985) and relative-order judgements of autobiographical episodes (Fuhrman & Wyer, 1988). It may follow that on a longer semantic list, such as the English alphabet, we should observe the same congruity effect.

On the other hand, there are reasons to expect a congruity effect may not be found with the English alphabet. Although temporally presented JOR lists are presented in a forward sequence, a backward directional search strategy is found on both subspan "later" instruction (Chan et al., 2009) and supraspan JOR lists (Chapter 2, Liu et al., 2014). However, alphabetical order is learned in the forward direction and is strongly forward-directional (Lovelace & Snodgrass, 1971; Grenzebach & McDonald, 1992; Klahr et al., 1983; Scharroo, Leeuwenberg, Stalmeier, & Vos, 1994). If long-list JORs are best understood as being carried out with a serial-recall-based strategy, it may be that participants only search in one direction (forward) for a list that was practised so many times in that forward direction; thus, if the congruity effect depends on, even partly, a reversal of search direction (Chan et al., 2009; Liu et al., 2014), that effect might simply not occur for over-learned, highly directional materials.

To test whether there is a congruity effect with the English alphabet, with both response-time and error-rate as measures, we tested the full range of the English alphabet, with all possible probe combinations (with equal probability), manipulating instruction between-subjects. We tested a large sample because we expected that even if response-time congruity effects were sizeable and easily measured, the error-rate effects might be quite subtle, given that accuracy was expected to be near-ceiling.

3.2 Methods

3.2.1 Participants

A total of 340 undergraduate students from introductory psychology courses at the University of Alberta participated in exchange for partial course credit. Participants gave informed consent, had normal or corrected-to-normal vision and learned English before age 6. We manipulated Instruction ("earlier," "later") between-subjects. Participants were run in groups

of about 10–15 with all participants within a testing group being assigned to a single experimental group; experimental group cycled across groups. Twelve participants were excluded because their accuracy was below 80%, or they self-reported having not followed the instructions. Final analyses thus included 173/175 and 155/165 participants in "earlier" and "later" groups, respectively. In Chapter 2, Experiment 2, (Liu et al., 2014), we collected 385 participants. Half of those were tested on LL=4 (episodic memory lists), for which accuracy was close to ceiling, but we failed to detect a congruity effect using the error rate measure. We therefore collected twice the sample size to increase power to detect a congruity effect in error rate on JORs of the alphabet, which was also expected to show near-ceiling accuracy.

3.2.2 Materials and Procedure

The experiment was created and run using the Python Experiment-Programming Library (Geller et al., 2007). Probes were pairs of the 650 possible permutations of the 26 letters of the English alphabet, in randomized presentation-order. Participants in the "earlier" group were asked to select which of the two probe letters comes earlier in the English alphabet. Participants in the "later" instruction group were asked to select which of the two probe letters comes later in the English alphabet. Participants were instructed to respond as quickly as possible without compromising accuracy. A single session lasted approximately one hour. The session started with a practice block of 8 trials, followed by 13 blocks of 50 trials. The computer provided immediate accuracy feedback after each trial in the practice block ("correct", "incorrect"), and average response time (ms) and accuracy (% correct) at the end of each experimental block. Each trial began with a fixation asterisk, '*', in the center of the screen, followed by the probe consisting two letters from the English alphabet, after which the participant made their response by pressing '.' key (for the left-hand probe item) or the '/' key (for the right-hand probe item). After a 500-ms delay, initiation of the next trial was self-paced. Trials with response time faster than 200 ms or slower than four standard deviations above a participant's mean response time were removed from the data (0.98% of all trials).

Linear mixed effects (LME) analysis (Baayen et al., 2008; Bates, 2005) was applied to our data to determine how instruction affected error rates and response time. LME was selected because it can fit individual responses without need for averaging the data, and protects against type II error due to increased power (Baayen et al., 2008; Baayen & Milin,

2010). LME analysis was conducted in R (Bates, 2005), using the lme4 (Bates & Sarkar, 2007), LanguageR (Baayen, 2007) and LMERConvenienceFunctions (Tremblay, 2013) libraries. The "lmer" function was used to fit the LME model. The "pamer.fnc" function was used to calculate the p values of model parameters. Instructions ("earlier," "later"), Serial position (serial position of the probe item that appeared earlier from the presented list), Distance (absolute value of the difference between two probes' serial positions), Intact/Reverse (whether the probe order was consistent/inconsistent with presentation order, respectively) and Trial number were included as fixed factors. Subject was included as a random factor. Instruction and Intact/Reverse were treated as categorical factors. All other factors were scaled and centered before being entered in the model. Response time was analyzed for correct trials only, and was log-transformed to reduce skewness. The error rate data were fitted with logistic regression as it is a binary variable ("correct" vs. "incorrect"). LME estimated random effects first, followed by fixed effects. In the results tables, the "Estimate" column reported the corresponding regression coefficients, along with their standard errors. For the purposes of reporting the LME results, the Intact condition and "earlier" instruction were set as the reference levels for the Intact/Reverse and Instruction factors, respectively. The best fits of LME models were obtained by conducting a series of iterative tests comparing progressively simpler models with more complex models using the Bayesian Information Criterion (BIC), was done in Chapter 2 (Liu et al., 2014), using LMERConvenienceFunctions (Tremblay, 2013).

3.3 Results

We first asked whether the instructions differed in difficulty by comparing the number of excluded participants (see methods) in each instruction (2/175 and 10/165, for "earlier" and "later" instructions, respectively). A chi-square test found differences between number of included subjects was significant ($\chi(1)^2$=35.32, $p < 0.001$). Thus, by this very coarse measure, the "later" instruction appears to be more challenging.

Next, we looked to the LME results to find out if the congruity effect was significant. The congruity effect is characterized by a crossover interaction between Instruction and Serial position. Serial position of the probe is defined as the serial position of the earlier probe. Note that defining serial position as the earlier probe serial position collapses across many

different Later probe values. However, there was a similar congruity effect when replotted as a function of the probe with higher serial rank (Figure 3.1 and Figure 3.2). The instruction by serial position interaction and "earlier" − "later" mean difference were plotted for response time (Figure 3.3) and error rates (Figure 3.4).

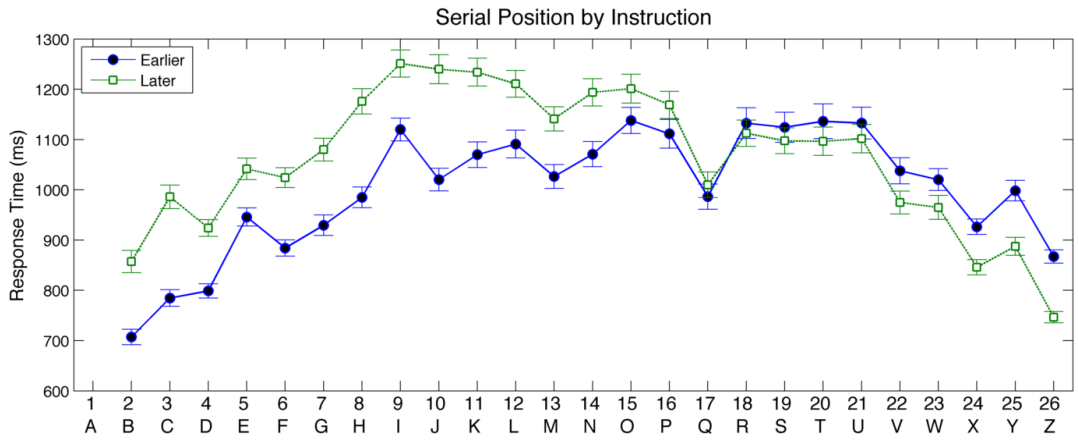
Response Time. The congruity effect can be clearly seen in response-time figure (Figure 3.3)— namely, the "earlier" instruction produced faster responses than the "later" instruction earlier in the alphabet and vice versa later in the alphabet, with a crossover point near serial-position 9 (the letter I). This was confirmed by a significant Instruction × Serial position interaction in the best-fitting LME model (Table 3.1).

		Estimate (SE)
Main effects		
	Intercept	6.815 (0.015)*
	Intact/Reverse	0.007 (0.003)*
	Instruction	-0.021 (0.022)*
	Trial	-0.045 (0.001)*
	Serial position	0.146 (0.002)*
	Distance	-0.098 (0.002)*
Interactions		
	Intact/Reverse × Instruction	0.0309 (0.003)*
	Intact/Reverse × Serial position	-0.069 (0.003)*
	Intact/Reverse × Distance	-0.045 (0.003)*
	Instruction × Trial	-0.002 (0.001)
	Instruction × Serial position	-0.135 (0.003)*
	Instruction × Distance	-0.065 (0.003)*
	Trial × Serial position	0.001(0.001)*
	Trial × Distance	0.004 (0.001)*
	Serial position × Distance	0.002 (0.002)*
	Intact/Reverse × Instruction × Serial position	-0.019 (0.004)*
	Intact/Reverse × Instruction × Distance	-0.022 (0.004)*
	Intact/Reverse × Serial position × Distance	-0.016 (0.002)*
	Instruction × Trial × Serial position	-0.009 (0.002)*
	Instruction × Serial position × Distance	-0.032 (0.002)*

Table 3.1: The best-fitting LME model for response time. The congruity effect is in bold. The "Estimate" column reports the corresponding regression coefficient, along with its SE (standard error). Significant effects are denoted * - $p < 0.05$.

The difference in mean response time between "earlier" and "later" instruction shows an

a



b

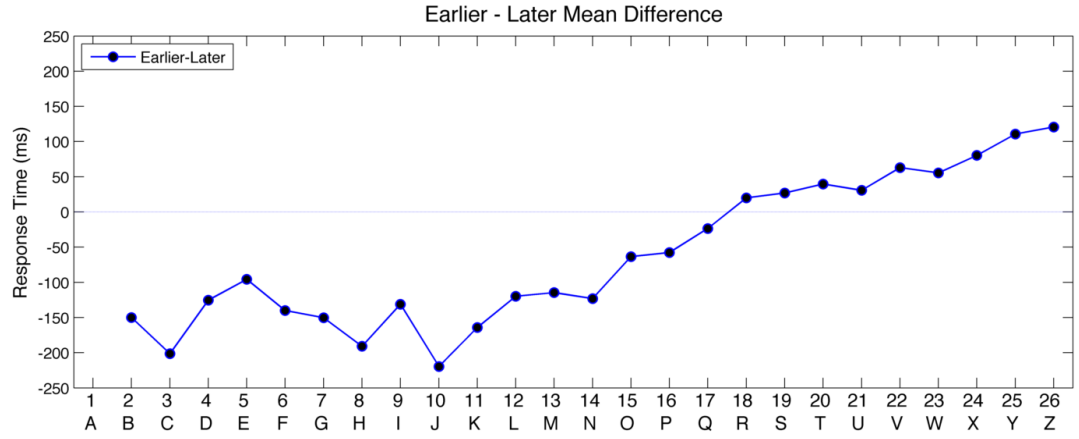
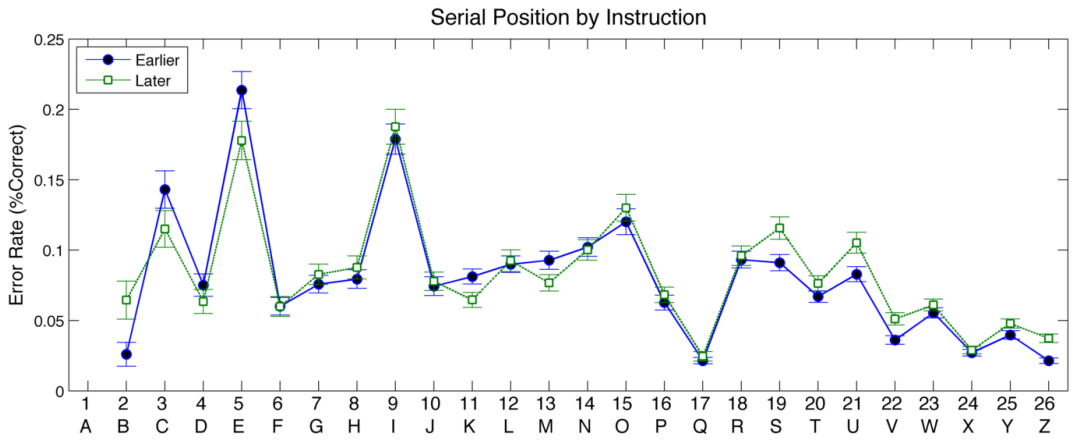


Figure 3.1: a) Instruction as a function of serial position with the response time measure. Serial position is defined as serial position of the **later** probe item. Error bars plot standard error of the mean. b) Earlier–later instruction mean differences.

a



b

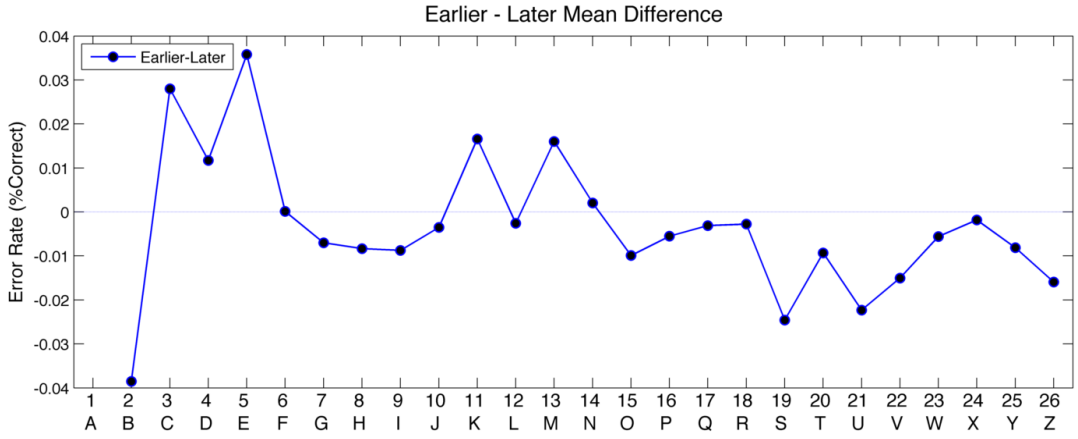
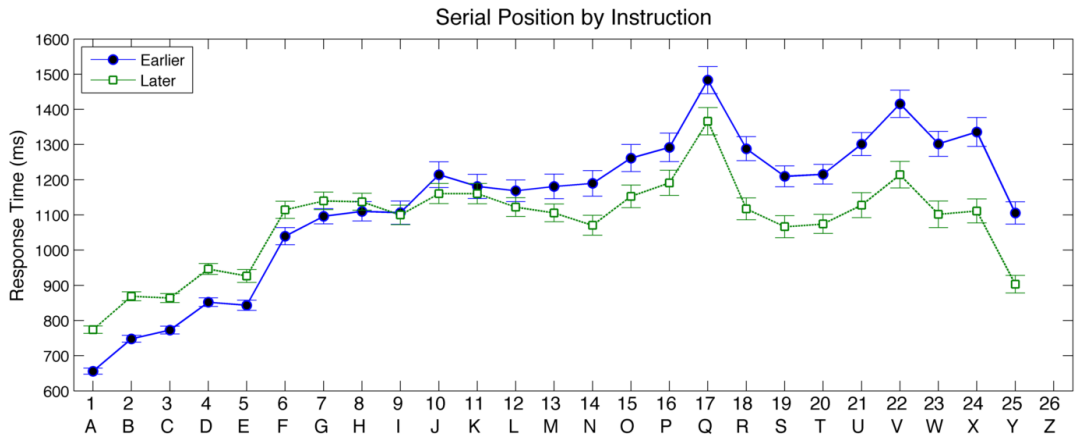


Figure 3.2: a) Instruction as a function of serial position with the error rate measure. Serial position is defined as serial position of the **later** probe item. Error bars plot standard error of the mean. b) Earlier–later instruction mean differences.

a



b

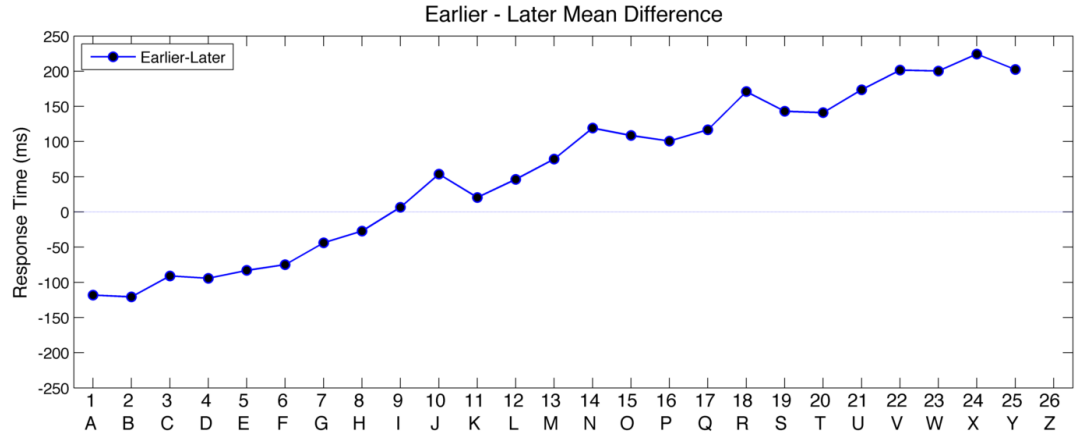
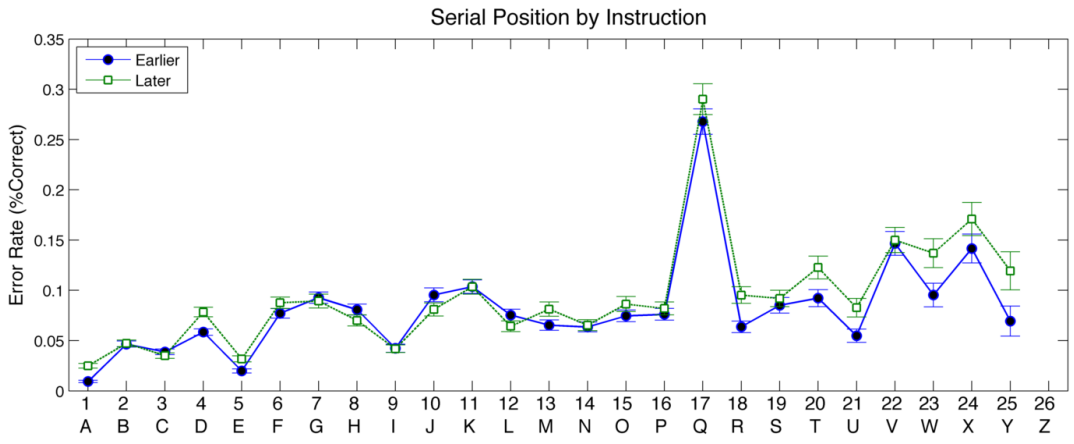


Figure 3.3: a) Instruction as a function of serial position with the response time measure. Serial position is defined as serial position of the **earlier** probe item. Error bars plot standard error of the mean. b) Earlier–later instruction mean differences.

a



b

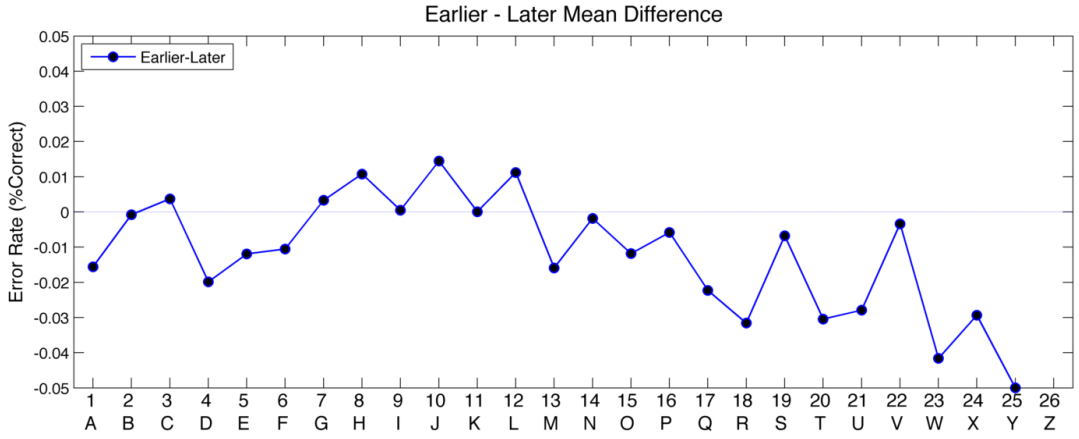


Figure 3.4: a) Instruction as a function of serial position with the error rate measure. Serial position is defined as serial position of the **earlier** probe item. Error bars plot standard error of the mean. b) Earlier–later instruction mean differences.

approximately linear increase from the earlier-probe serial position 1 to 25, as can also be seen in the interaction plot generated by the best-fitting LME mode (see Figure 3.5a). For Instruction = "earlier", response time is expected to be higher as serial position increased at a rate of 0.146 log-transformed milliseconds per Serial position, but for Instruction = "later" the effect of Serial position is -0.135 unit lower. Note that the congruity effect is of the same order of magnitude as the serial position effect, one of the most robust findings in serial-order memory research.

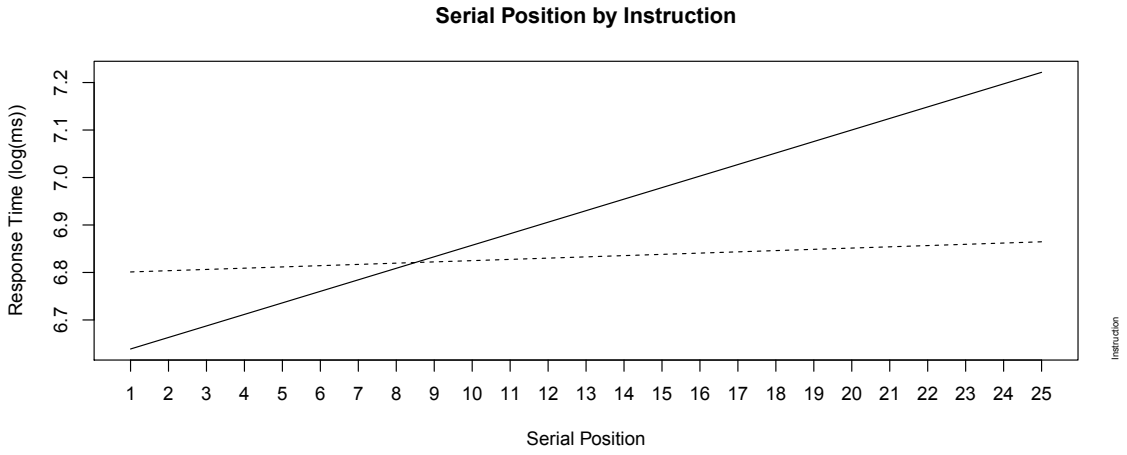
The Instruction × Serial position interaction was also part of higher-order interactions, including Intact/Reverse, Distance and Trial. However, in all these higher-order interactions, the form of the two-way interaction was the same (see Figure 3.6), and therefore do not speak to our main objective, to test for a congruity effect.

Error rate. Turning to the error-rate measure, inspection of the plot of Instruction × Serial positions (Figure 3.4) suggests no congruity effect. The best-fitting LME model for error rates also failed to find a significant Instruction × Serial position interaction. Comparing the error-rate data (Figure 3.4) to the response-time data (Figure 3.3) suggests that "earlier"−"later" instruction difference at higher serial positions shows an opposite pattern for response time and error rates, showing a speed-accuracy tradeoffs.

To follow up further on the speed-accuracy tradeoff, we tested our visual impression that the speed-accuracy tradeoff mainly occurs at higher serial positions. We plotted response time against error rate to find out if they were correlated, but we divided the alphabet into two halves, with the first half from serial position 1 to 13 (A to M) and second half from serial position 14 to 25 (N to Y). Figure 3.7 shows the relationship between error rate and (correct-trial) response times for the difference between instructions, broken down into three conditions: a) both probes within the first half of the list, b) both probes within the second half of the list, and c) probes in different halves of the list. Pearson correlations between error rates and response time across probes were not significant within the first half of the alphabet, $r(328) = 0.012,\ p > 0.5$, nor for probes that crossed between the first and second halves of the alphabet, $r(328) = 0.031,\ p > 0.5$; but it was significant for probes within the second half of the alphabet, $r(328) = -0.294,\ p < 0.01$.

This led us to suspect that the congruity effect in response time might be rushing participants too much toward the end of the alphabet in "later" instruction, with the side-effect of producing more errors late in the alphabet in that group. In other words, many of the

a



b

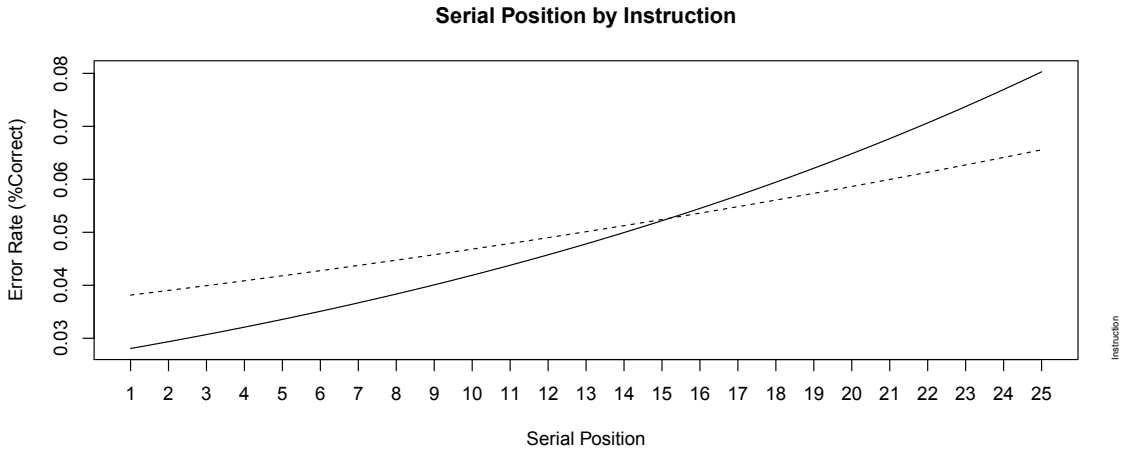
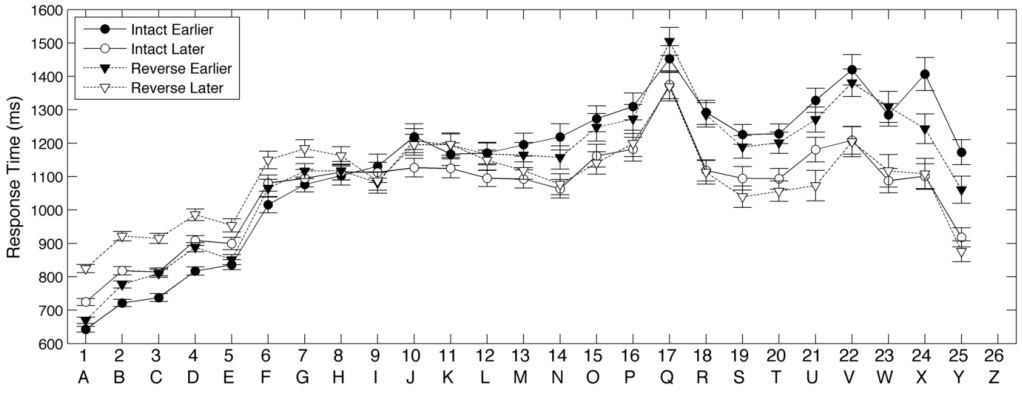
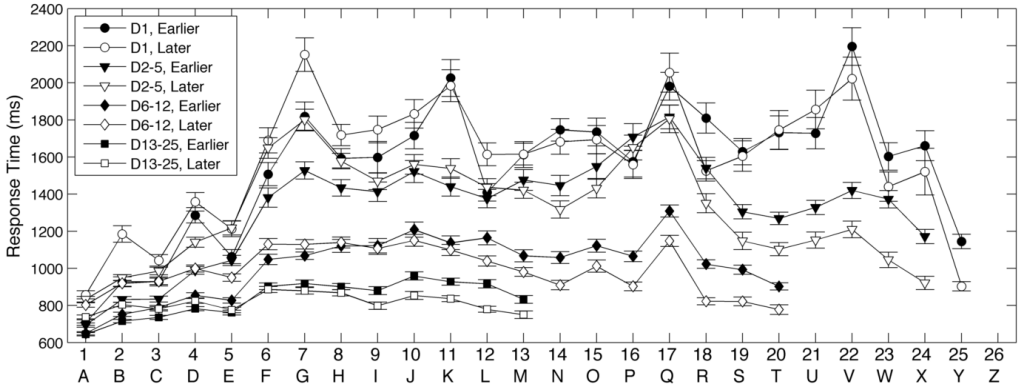


Figure 3.5: Instruction by serial position interaction generated from the best-fitting LME model for a) log-transformed response time and b) error rates with response time entered as a predicting factor. The solid line represents "earlier" instruction and dashed line represents "Later" instruction. Serial position is defined as serial position of the earlier probe item.

a



b



c

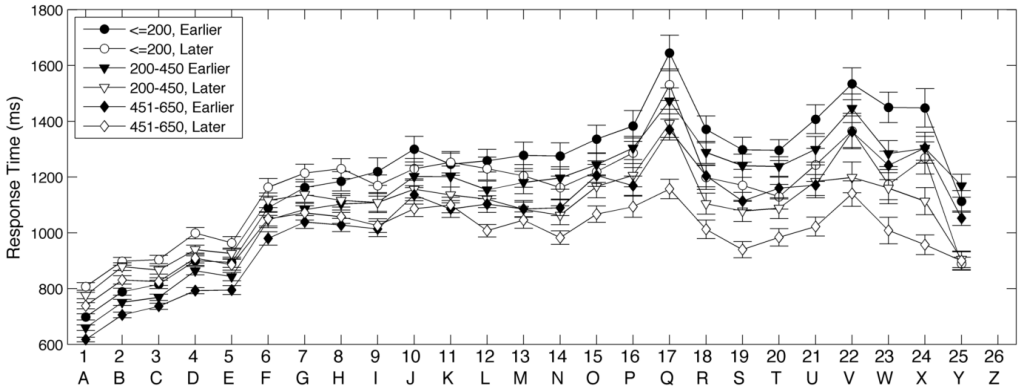


Figure 3.6: a) Each pairwise combination of instruction and Intact/Reverse plotted as a function of serial position. b) Instruction and binned Distance plotted as a function of serial position. c) Instruction and binned Trial numbers plotted as a function of serial position. Serial position is defined as serial position of the earlier probe item. Error bars plot standard error of the mean.

		Estimate (SE)
Main effects		
	Intercept	2.681 (0.167)*
	Intact/Reverse	0.275 (0.024)*
	Instruction	0.167 (0.070)*
	Serial position	0.272 (0.024)*
	Distance	-1.051 (0.026)*
	Response Time	-0.898 (0.024)*
Interactions		
	Intact/Reverse × Serial position	-0.350 (0.017)*
	Intact/Reverse × Distance	-0.392 (0.028)*
	Instruction × Serial position	-0.158 (0.030)*
	Instruction × Distance	0.157 (0.028)*
	Serial position × Distance	-0.020 (0.021)*
	Instruction × Serial position × Distance	-0.140 (0.029)*

Table 3.2: The best-fitting LME model for error rates. The congruity effect is in bold. The "Estimate" column reports the corresponding regression coefficient, along with its SE (standard error). Significant effects are denoted * - $p < 0.05$.

faster responses may have been errors, and more so for "later" group than "earlier" group. Therefore, if speed were controlled for (on both correct and error trials), we could then test whether error rate showed a underlying congruity effect. We therefore controlled for the speed-accuracy tradeoff by adding response time as a fixed effect to the best-fitting LME model for error rate (Table 3.2). We found that increased response-time was associated with less error, confirming the general presence of a speed-accuracy tradeoff. The best-fitting model then found a significant Instruction × Serial position interaction, which can be visualized in the model-produced interaction plot (Figure 3.5b). For Instruction = "earlier", error rate is expected to be higher as serial position increased at a rate of 0.272 logit per Serial position, but for Instruction = "later" the effect of Serial position is 0.158 unit lower. In line with the response time results, the congruity effect is of the same order of magnitude as the serial position effect. The error-rate congruity effect produced by the best-fitting LME model is consistent with the congruity effect found in the (correct-trial) response-time analysis. The two-way interaction is qualified by a Instruction × Serial position × Distance three-way interaction (see Figure 3.8), in that larger Distance and Serial position were associated with less error when the Instruction was "later" than "earlier," but those higher-order

interactions did not indicate a change in the form of the basic congruity effect.

3.4 Discussion

We found a congruity effect in a very long, semantic list (the English alphabet), with response time as the measure. A congruity effect was also found with error rate as the measure, after controlling for speed-accuracy tradeoffs. This builds on comparable findings with subspan (Chan et al., 2009) and supraspan (Chapter 2, Liu et al., 2014) temporal-order JOR procedures.

The congruity effect may therefore be a benchmark phenomenon for order-memory judgements, shared across list length and list property (temporally presented lists versus semantic lists). To our knowledge, models of serial-order memory have not been designed to produce congruity effects, and thus, the congruity effect may be diagnostic of memory models or suggest how existing models might be further developed (Chapter 2, Liu et al., 2014). Theories designed to explain comparative judgements may also shed light on our understanding of congruity effects across cognitive domains.

Our confirmation of a congruity effect using the 2AFC procedure suggest participants may voluntarily chose to use different strategies for the "earlier" and "later" instruction, and this finding may mean that Jou and Aldridge's (1999) prior failure to find a congruity effect with the 2AFC procedure was due to insufficient sample size. Jou found only a a main effect of instruction, and no interaction between instruction and serial-position, in JORs of the first 9 letters of English alphabet. The congruity effect is expected when participants are asked to make responses from a continuum, for the English alphabet the range is 26 items from A to Z. Hinrichs (1970) found that absolute order judgements scaled with the perceived maximum list size. Thus, the congruity effect might also be dependent on the perceived set size. Jou limited the probes to the first 9 letters from the alphabet, but participants may still have conceptualized the alphabet as a list of 26 items. In this case "earlier" instruction is congruent with the first 9 letters of the alphabet because they are in the first half of the 26 range, but the later Instruction is not particularly congruent with any of the items tested. Or, in more subtle form, the congruity effect may be easy to observe across the whole list, but may become quite subtle at the edges. To better understand this result, we filtered our data set to include only probes from the first 9 letters of the alphabet and also

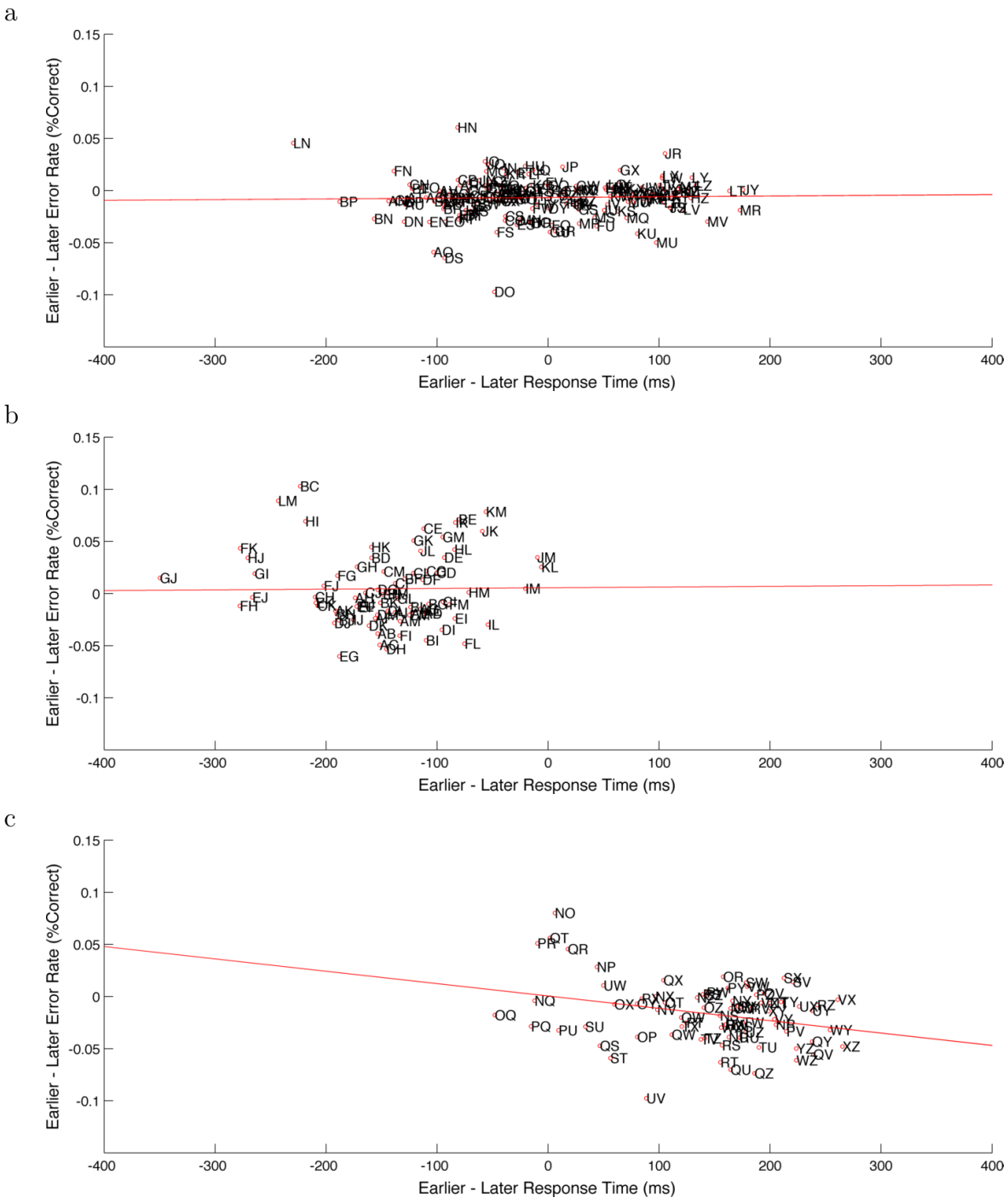
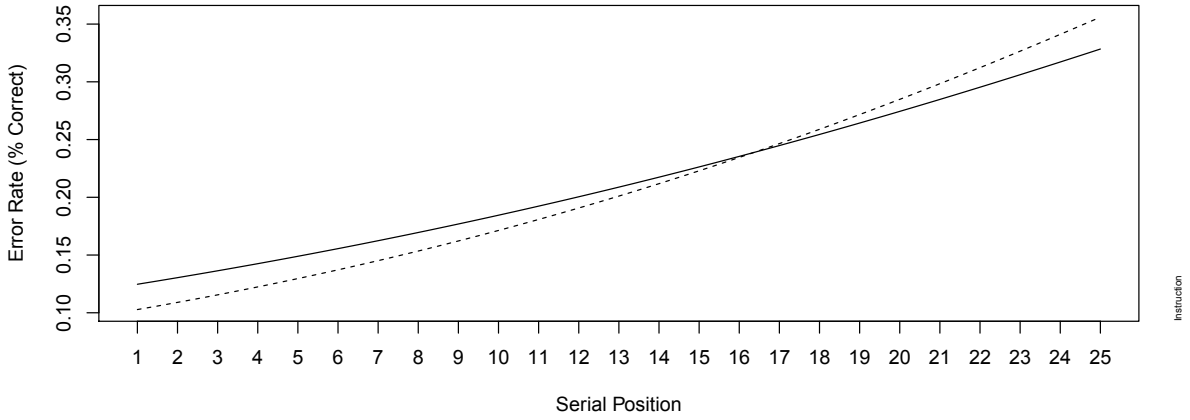
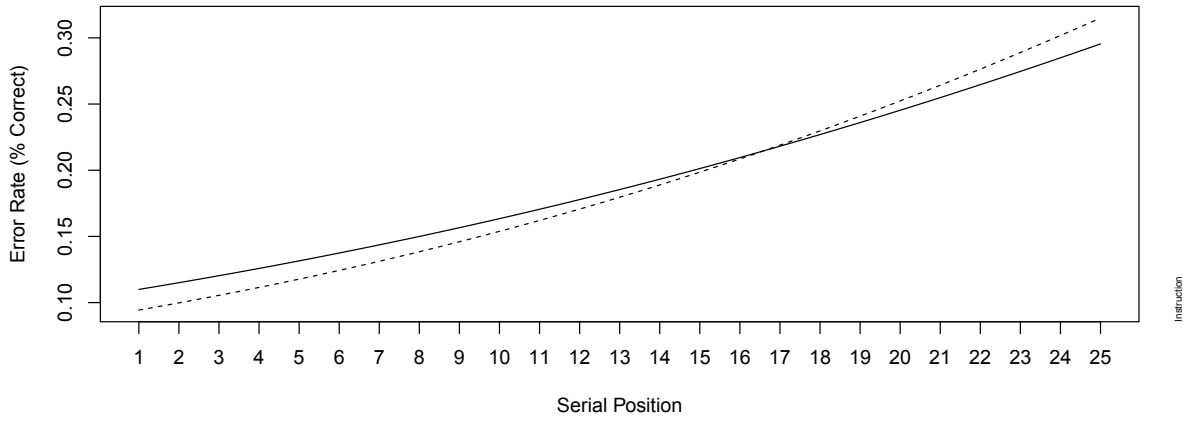


Figure 3.7: ER and RT differences for each probe pairs, averaged across participants: a) Probes across first and second half of the alphabet list ($y = 0.000007 * x - 0.0067$); b) probes within the first half of the alphabet list ($y = 0.000007 * x - 0.0055$); c) probes within the second half of the alphabet ($y = 0.00012 * x + 0.00054$)

a



b



c

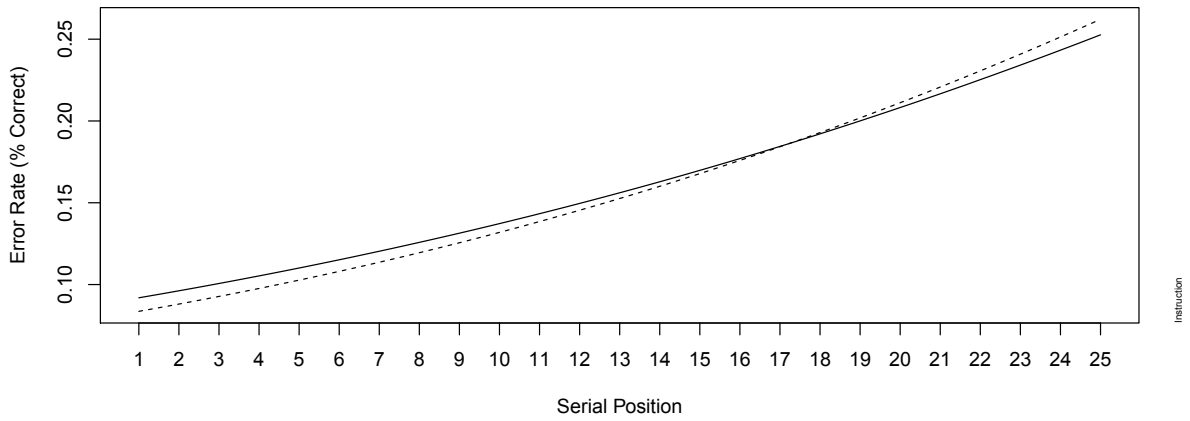


Figure 3.8: Instruction by Serial position interaction generated from the best-fitting LME model for Distance 1, 5 and 13 (panel a, b, c respectively). Serial position is defined as serial position of the earlier probe item.

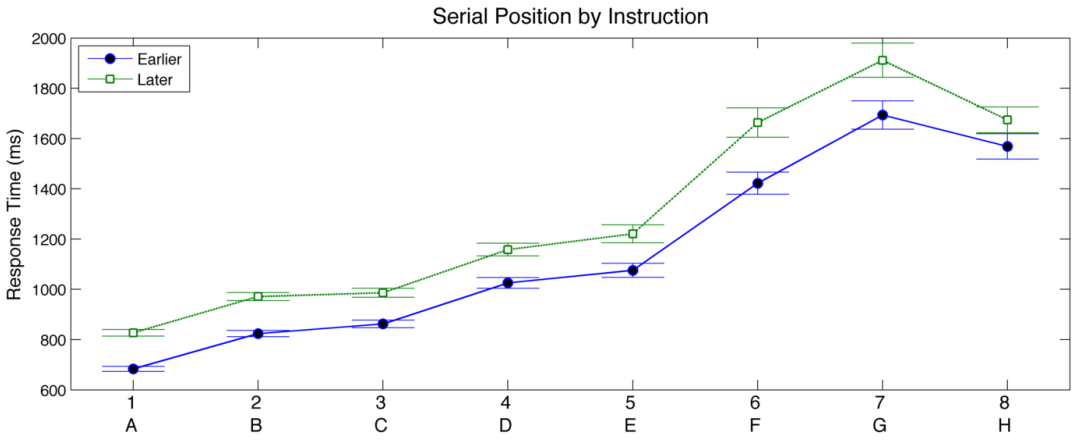


Figure 3.9: Instruction by serial position interaction as a function of response time, when both probes were from the first 9 letters of the English alphabet. Error bars plot standard error of the mean.

found no significant Instruction × Serial position interaction (Figure 3.9), replicating this null interaction. We also replicated a main effect of Instruction (faster response time for the earlier instruction), which is in line with Jou's (2003) results if a set of letters from the alphabet is always perceived as a set of 26 items, regardless of the JOR testing range.

In addition to the response-time congruity effect, we found an error-rate congruity effect after controlling for speed-accuracy tradeoffs. An error-rate congruity effect, to our knowledge, has not previously been reported for the English alphabet. Speed was not significantly associated with accuracy for the first half of the English alphabet; however. faster response time was significantly associated with lower accuracy for the later half of the alphabet, opposite to supraspan JOR results in Chapter 2 (Liu et al., 2014). In contrast to the robust finding of the response time congruity effect, error-rate congruity effects have only rarely been found in comparative judgements. One possible explanation is that error-rate congruity effect is very subtle and a large N and large number of trials per condition per subject is required to detect this effect. In addition, with tasks involving materials that are overlearned, researchers, understandably, typically focus on response-time as the principal behavioral measure (e.g., Birnbaum & Jou, 1990). Petrusic (1992) demonstrated an error-rate congruity effect on a very complex perceptual task, where two pairs of dots were presented and the judgement is based on selecting the pair of dots closer to a horizontal or vertical line.

The finding of a speed-accuracy tradeoff of the instruction differences only for the second half of the list challenges most existing accounts of the congruity effect from the comparative judgements literature. In reference point theory (Holyoak, 1978; Jamieson & Petrusic, 1975; Marks, 1972), each judgement is made by comparing both stimuli to a reference point. The ratio of differences between stimuli and reference point can thus account for the congruity effect. However, reference point theory would predict response time and error rate to be consistent in direction, and would not be compatible with error rate and response time effect opposite in direction. Semantic coding theory (Banks, 1977) is also a classic model of the semantic congruity effect. The model assumes a semantic code is first generated (e.g., "earlier", "later") and the code can be directly compared with each other. A decision can be made if the code differs. If both codes are the same, further processing is required to further refine the semantic code. For instance, when B and C are both coded as "early" from the alphabet, a further refinement has to be made to translate the code of B to be "very early" to make a decision. The semantic congruity effect is explained by assuming the instruction congruent with the semantic code would require no code translation, but the incongruent condition would. Similar to reference point theory, no speed-accuracy tradeoff is expected with this model. Birnbaum and Jou (1990) proposed the congruity effect could be explained by differential bias caused by the instruction, implemented in a random-walk-based model. The model would also inevitably predict the speed and accuracy pattern should be consistent.

One possible explanation of the speed-accuracy tradeoff may be a guessing strategy moderated by subjects' willingness to make a guess. Although it does not explicitly model serial-position effects, Petrusic's (1992) Slow-Fast Guessing Theory (SFGT) assumes information accumulates in counters and a decision can be made when a threshold is reached for any counter. This is different from a random-walk-based model (e.g., Ratcliff, 1978; Link, 1990) where information favoring opposite decisions cancels out and is represented by a single drift rate. SFGT predicts differential error rate and response time pattern by assuming three separate counters collect evidence favoring either probe and evidence favoring indecision. Indecision evidence is when evidence accumulated favors neither probe, therefore a guess is required. There is a fixed probability that evidence will accumulate in any counter per trial of evidence accrual. A decision is made when a threshold is reached for the counter favoring one decision before the threshold for indecision is reached, or else a guess is made. In this

model, response time is a function of amount of evidence accumulated, error rate is determined by the threshold for each counter and evidence accumulated in each counter. When the threshold is higher for a guess, and the counter for indecision has a lower probability of increasing, a slower response time and less error is predicted. The opposite prediction, faster response time and more error, is predicted when indecision evidence is more likely accumulated and the threshold to make a guess is low. If the first half of the alphabet is better encoded compared with the second half of the alphabet, we can assume order judgements in the second half of the alphabet requires more guessing. The SFGT model might help explain our data, by assuming the threshold to make a guess is lower for for the later serial positions.

As we have discussed previously, congruity effects already require extensions to current models of order-memory. Moreover, as with comparative judgement theories, speed-accuracy tradeoffs also challenge memory models. Many memory models are not developed to explain both speed and accuracy data (e.g., Brown et al., 2007; Lewandowsky & Murdock, 1989). Hacker's self-terminating search model (Hacker, 1980) handles both response time and error rate; however, response time is derived from error rate directly, meaning that speed and accuracy cannot trade off. OSCillator-based Associative Recall (OSCAR; Brown et al., 2000) is another model that has been fit successfully to JOR data. OSCAR assumes items are associated with the state of a internal context signal, and retrieving items requires reinstatement of that context. In the JOR task, the end-of-list context vector is used as a probe and the strongest activated list item is compared to the probe. If a match is found the search terminates. If a match is not found, the search continues to the next highly activated item. OSCAR predicts response time by the overall number of comparisons, which allows the response time pattern to deviate from the error rate pattern. However, there is no obvious explanation of how this mechanism could explain a speed-accuracy tradeoff, nor how it could handle a congruity effect, given that the response time data showed both primacy and recency effect, and the error rates are dominated by an overall primacy effect.

3.5 Conclusion

In conclusion, our findings of congruity effects on alphabetical order judgements provide further evidence that the congruity effect is a general phenomenon found not only in episodic

(temporally ordered) lists, but also in long, semantic lists. This strengthens the argument that memory judgements of order may be best thought of as a subset of comparative judgements. This also implies that congruity effects on comparative judgements may also materialize in error rate data, with sufficient power. Finally, the finding of a speed-accuracy tradeoff in the congruity effect that was specific to the last half of the alphabet presents a challenge to current models of serial-order memory as well as models of comparative judgements.

Chapter 4

Effects of grouping on forward and backward serial recall

4.1 Introduction

Order memory is essential for our daily functions (Lashley, 1951). One way of enhancing the overall order memory accuracy is to organize information into groups (e.g., Miller, 1956; Wickelgren, 1967; Ryan, 1969a). Groups are mini-lists that provide substructure for the full list. For example, a 10-digit telephone number "17804271432" could be organized into four groups, "1" is the country code, "780" is the area code, "427" is the exchange and "1432" is the local number. Grouping the 10-digit phone number into country code, area code, exchange code and number may enhance the overall recall accuracy, and this enhancement of memory maybe at the expense of mixing up numbers within each group (e.g., "1432" may be mistakenly remembered as "1423") or between groups (e.g., "14277801432", swapping area code "780" and exchange code "427").

Grouping has been extensively studied in the serial recall paradigm, where participants recall a studied list in order. It has been found that serial position curves show non-smooth accuracy transitions between certain adjacent positions, and this non-smooth characteristic has been used to suggest that people group lists spontaneously (e.g., Martin & Noreen, 1974; Madigan, 1980; Jou, 2011). Apart from the spontaneous grouping behaviour, grouping can be induced by a broad spectrum of laboratory methods by adding cues about group boundaries, such as temporal pauses (e.g., Hitch, Burgess, Towse, & Culpin, 1996; Frankish, 1985; Maybery et al., 2002), spatial grouping (e.g., J. A. Anderson, Silverstein, Ritz, & Jones, 1998; Parmentier & Maybery, 2008), voice cues (Frankish, 1989; Parmentier & Maybery,

2008), and overt instruction to form groups (e.g., Farrell, 2012a). The effects of grouping on forward serial recall have been studied extensively (for a review see Terrace, 2001), and this has led to several hypotheses about how grouping influences serial-order memory. However, as we explain below, backward recall of grouped lists may provide additional evidence that can help us to decide amongst the alternative accounts of grouping. To date, the effects of grouping on backward serial recall have received little attention. Next we briefly review theories of grouping derived from the forward serial-recall results, and then consider how backward serial recall could contribute to this discussion.

In forward recall, grouping enhances the overall recall accuracy, and, more specifically, induces a "scalloped" pattern in serial position curves, which may be evidence of grouping. (Henson, 1998; Lee & Estes, 1981; Maybery et al., 2002; Ng & Maybery, 2002, 2005; Wickelgren, 1967; Ryan, 1969a). The "scalloped" pattern is characterized by a mini serial position curve for each group, each showing within-group primacy and recency effects. In addition, grouping affects the kinds of errors participants make in very specific ways (Ryan, 1969a, 1969b). Order errors are generally defined as when the participant responds with a list item, but in the wrong output position. A transposition error is when two items swap positions. The most common kind of transposition error is adjacent transpositions (Lee & Estes, 1977), that is when an item is remembered at an adjacent serial position. Compared to ungrouped lists, adjacent transposition errors are lower overall and especially lower when two items are across a group boundary. In exchange, grouping introduces another kind of transposition error called interposition error, that is when an item is recalled in the wrong group but with the correct within-group position (Henson et al., 1996; Farrell & Lewandowsky, 2004). For example, for a list of 9 items comprised of three groups: 3-3-3, item 2 being recalled in position 3 would be an adjacent transposition error, whereas item 2 being recalled in position 5 would be an interposition error.

Latency data also reveal how participants remember grouped lists. Participants are slower to recall the first item within each group, followed by much faster latency for the remaining within-group items (Maybery et al., 2002; Ng & Maybery, 2002, 2005; Farrell & Lelièvre, 2012; Thomas et al., 2003). Thus, recall is faster within-group than between-group. The latency data provide complementary evidence that the grouping manipulation of the experiments are effective.

The accuracy and latency results for forward serial recall have been taken as support for

the idea that people use a positional code as a retrieval-cue in serial recall, where item position can be represented relative to the whole list, or relative to each group (J. R. Anderson & Matessa, 1997; Hurlstone et al., 2014). Taking a 9-item list composed with groups of 3 items as an example, the position codes for the whole list is 123456789, the position code within each group is 123 (i.e., the whole list represented by within-group position is 123123123). An alternative to the list-level position code could be the position code of the individual items (Brown et al., 2000, 2007; Burgess & Hitch, 1999; Lewandowsky & Farrell, 2008) (e.g., positional coding: 123456789), or the position of groups (J. R. Anderson & Matessa, 1997; Farrell, 2012b; Henson, 1998; Lee & Estes, 1981) (e.g., positional coding: 111222333).

According to the hierarchical positional-coding account, ungrouped lists require a single dimension of positional codes, whereas grouped lists require two levels of position codes. Consequently, position is coded at a higher level of precision in grouped lists. The enhanced level of precision from hierarchical positional coding is used to explain why grouped lists are overall more accurate than ungrouped lists. One sub-class of position codes are temporal codes, where the position codes are not just ordinal, but also coded by the presentation times (Brown et al., 2000, 2007). Ordinal position codes and temporal codes can be represented as separate dimensions (Brown et al., 2007). The positional- or temporal-distinctiveness accounts of grouping predict a specific pattern of serial-position effects; namely, in a grouped list, the items at the start and end of each group are more distinct than they would be in ungrouped lists, so they should be more competitive and be recalled better. The middle items of groups are less distinct, so they should be at a relative disadvantage. Although there is considerable support for these predictions, Parmentier, King, and Dennis (2006) have criticized positional-distinctiveness accounts of grouping, and we shall discuss their argument in the Discussion.

Positional coding models often implicitly assume groups are processed in the forward direction (e.g., J. R. Anderson & Matessa, 1997; Farrell, 2012b). Recall of the last few items as a forward-ordered group is found in free recall (e.g., Beaman & Morton, 2000; Bhatarah, Ward, & Tan, 2008; Farrell, 2010; Grenfell-Essam & Ward, 2012), which supports the idea that groups are forward-coded. The forward positional coding assumption predicts that backward recall may require extra resources to reverse the order of forward-coded positions at retrieval, thus predicting the backward recall to be slower, overall, than forward recall, and the forward recall is predicted to be more accurate, or as accurate as backward recall.

Alternatively, if we assume position codes could be processed in backward direction without additional effort, the output position curve, when the behavioural measure is aligned using the output position, for forward and backward recall should be identical, and the serial position curve for forward and backward recall should be mirror images.

Associative chaining versus positional coding

An alternative perspective to explain serial order memory is the concept of associative chaining, where serial order memory is formed by item-to-item associations. For example, the list ABCDE could be learned by learning associations of AB, BC, CD and DE. Then, A is used as a cue to recall B, and the just-recalled B is used as a cue to recall C, the already recalled items are excluded from the pool of available response candidates to prevent repeated items. The process continues until the full list is recalled in sequence. The Theory of Distributed Associative Memory (TODAM) (Lewandowsky & Murdock, 1989) is a model that has been used to simulate the associative chaining process, by representing items with random vectors and representing associations by convolution of item vectors. Convolution is a mathematical operation to combine two item vectors where the original item vectors could be retrieved by the mathematical operation correlation (see Lewandowsky & Murdock, 1989). TODAM stores memory in a common vector, where items can be retrieved by item associations. An important property of TODAM is that the convolution operation is commutative. That means that associations are remembered without order (A–B and B–A are stored identically in TODAM). Thus, TODAM predicts associative symmetry; e.g., given a pair A–B, recalling A given B and should be equally accurate as recalling B given A. Associative symmetry has been confirmed in memory for associations (e.g., Kahana, 2002). In cued recall of serial lists, associative symmetry was either confirmed (Caplan, Glabolt, & McIntosh, 2006), or else only small-magnitude forward-probe advantages were found (Kahana & Caplan, 2002), suggesting that backward associations are at least nearly as strong as forward associations. Thus, at least without further assumptions, the associative symmetry property of TODAM predicts that the backward recall serial position curve should be very similar to forward serial recall. That is, instead of starting to recall the list cueing with the first item, backward recall can be straight-forwardly implemented by initiating recall with the last list item. In other words, the associative-chaining account of serial recall would treat backward recall as a direct readout of the list in the backward direction. More complex

extensions of simple chaining have been proposed. In one variant of TODAM, the power-set model (Murdock, 1995), remote associations between items also contribute to recall. In this model, the association of the current item with previous items are normalized, and the normalization factor differ between recall direction, predicting dissociations in forward and backward recall. However, simulations showed that forward and backward recall generated by the power-set model were almost mirror images (Murdock, 1995). This is partly because adjacent inter-item associations dominated the effects of remote associations.

To further examine the positional coding and associative chaining account of serial recall, both the accuracy and latency measures are important. Accuracy results are usually plotted as a function of serial position— i.e., position within the original presentation order. However, the latency measure, inter-response time, is calculated by the differences between onset of adjacent responses; thus it is more appropriate to plot response times as a function of output order (Thomas et al., 2003). J. A. Anderson et al. (1998) found that based on input order, the serial position effects of accuracy and latency both were mirror images between forward and backward recall. The accuracy and latency findings are compatible with positional coding and associative chaining models that assume backward recall access the list directly in reverse order. However, Thomas et al. (2003) found for subspan ungrouped lists, forward recall aligned by the output order showed a linear decrease in latency whereas backward recall aligned by output order showed a bow-shaped output position function. This finding is problematic for the idea that lists could be readout in backward order directly, inconsistent with assumptions of associative chaining models. Thus, it is not well established whether output position plots for both recall directions are similar or different. If the output position curves are similar between forward and backward recall, positional coding models can no longer assume groups are coded and retrieved in forward direction. Associative chaining models could be adapted to account for the direct readout of the list in backward order, as a reverse retrieval of the chain predicts the same output pattern between recall directions. Because serial position curves and output position curves provide complementary evidence that speaks to models, we present both.

The scan-and-drop strategy

Some researchers have argued against backward recall being a direct backward readout of a serial list, independent of the chaining versus positional-coding debate. Conrad (1965)

proposed a scan-and-drop strategy for backward recall, where participants perform multiple forward recalls starting from the beginning of the list, and "drop" the target item from the list when it has been recalled. In other words, the assumption is that the last item read-out is the item the participant recalls, and the recalled items are dropped from the list. For example, for a list with items A B C D and the backward recall task, the last list-item D is recalled by a forward scan of ABCD, C is recalled by a forward scan of ABC, B is recalled by a forward scan of AB and this process continues until the full list is recalled, thus predicting the longest recall time to recall D followed by a faster latency as output position increases. Thomas et al. (2003) found that backward serial recall latency on word lists (list length = 4, 5 and 6) declines monotonically after the second item in the list. This monotonic decline in latency was taken as evidence to support a scan-and-drop strategy.

However, evidence for scan-and-drop is mixed. Although Thomas et al. (2003) interpreted their findings as supporting scan-and-drop strategy, and the list length 4 latency pattern is consistent with scan-and-drop predictions, the face-value serial position effects did not seem to support scan-and-drop for list length 5 and 6. Thomas et al. (2003) found the peak latency for list length 5 and 6 of backward serial recall was at output position 2 for Experiment 1 and at output position 3 for Experiment 2. They argue this pattern could be explained by participants first outputting the last item in the list because it is immediately available, and then performing the scan-and-drop starting from the beginning of the list. Their explanation rests on the assumption that backward recall involves breaking down the list into two parts, outputting immediately available items first, followed by an elaborative scan-and-drop strategy. Although possible, this assumption may be difficult to test directly, and other accounts of their serial-position effects may be equally tenable. Haberlandt et al. (2005) used a verbal recall task in contrast to Thomas et al. (2003) and their data is also consistent with scan-and-drop strategy at list length 4, but required the same assumption as Thomas et al. (2003) to explain their list-length 5 and list-length 6 data. We speculate that the scan-and-drop strategy might be specific to short lists (e.g., in Thomas et al. (2003) Experiment 2, list length 6, forward recall showed a latency peak at output position 5, suggesting the first four items and last two items are processed as two groups). One of our goals is to test Thomas et al.'s (2003) account, whereby grouping complicates a scan-and-drop strategy, with our experiment.

The scan-and-drop strategy suggested by Thomas et al. (2003) and Haberlandt et al.

(2005) would predict forward and backward recall patterns are different in specific ways. The scan-and-drop strategy predicts an enhanced primacy effect on the whole list level and also at the group-level, because the earlier items within list or within each group receive more rehearsals. With the list ABCD as a example, the scan-and-drop process is summarized as: covertly retrieve ABCD - output D, covertly retrieve ABC - output C, covertly retrieve AB - output B, covertly retrieve A - output A. A is practiced 3 times before it is recalled, B is practiced twice before it is recalled. The simplest results would be backward recall showing a monotonic decline of latency as output position increases. If we further assume that scan-and-drop could operate at the level of each group, we would expect a monotonic decline of latency as within-group position increases.

Backward recall of grouped lists

Researchers have tried to find out whether backward serial recall operated on the same principles as forward serial recall, and tested benchmark effects from forward serial recall such as phonological similarity, word length, articulatory suppression and irrelevant speech using the backward serial recall paradigm (Bireta et al., 2010; Guérard, Saint-Aubin, Burns, & Chamberland, 2012). The results have been mixed, with some found all of the above mentioned benchmark effects generalize to backward serial recall (Guérard et al., 2012), and others found the word length effect does not generalize to backward serial recall (Baker, Tehan, & Tehan, 2012; Bireta et al., 2010; Surprenant et al., 2011), or all of the above mentioned benchmark effect does not generalize to backward serial recall (Bireta et al., 2010). Ritchie, Tolan, Tehan, and Goh (2015) conducted a meta analysis to explore the conflict findings between Bireta et al. (2010) and Guérard et al. (2012). Data from 16 experiments was re-analyzed by a technique of grouping the output positions into two groups, a output primacy group with the first two outputted items, and a output recency group with the other items (St. Clair-Thompson & Allen, 2013). Ritchie et al. (2015) found that the benchmark effects generalized to backward recall, but the effect is attenuated for the first two outputted items. We therefore wondered if the standard effects of grouping would replicate in backward recall. This could tell us whether grouping depends on serial position or output position, and whether grouping effects might be sensitive to some of the ways in which backward recall has been suggested to differ from forward recall. The simplest result we might obtain is that temporal grouping facilitates forward serial recall but not backward recall. This predicts the

serial position curve is identical for both ungrouped and grouped backward serial recall.

J. A. Anderson et al. (1998) conducted a closely related study investigated both forward and backward recall of grouped and ungrouped lists. They found the latency measure showed long pauses across group boundaries, and the effect of grouping was similar regardless of the recall directions. However, their lists were presented sequentially with groups denoted visually, both at study and test. In fact, serial position (of items within the whole list) were also denoted spatially both at study and test. Thus, their outcome could have been due to explicit position-cueing at study and test. This led us to predict that the same might be true of the temporal manipulations of grouping which have driven thinking about hierarchical positional coding in serial recall. Li and Lewandowsky (1993, 1995) and St. Clair-Thompson and Allen (2013) published evidence that backward recall is more likely to rely on spatial strategies than forward recall. The spatial demarcation used in J. A. Anderson et al. (1998) may have induced participants to rely more on spatial information for both recall directions, leaving open the possibility that a standard manipulation of grouping via temporal grouping might not show the same effects on backward as in forward recall.

In sum, backward recall may further our understanding of how grouping interacts with order memory, and may help differentiate theoretical accounts, such as the scan-and-drop strategy, positional coding and temporal-positional coding, and associative chaining models.

To study effects of temporal grouping and recall direction, we tested both forward and backward recall direction and both temporally grouped and ungrouped lists, using a between-subject design (Direction ["forward"/"backward"] × Grouping ["grouped", "ungrouped"). The "ungrouped" condition presents 9 consonants with identical inter-item intervals. A group size of 3 is thought to produce optimal performance and is commonly used in grouping studies (e.g., J. A. Anderson et al., 1998; Henson et al., 1996; Henson, 1999; Ryan, 1969a; Wickelgren, 1967); thus, the "grouped" condition presents 9 consonants broken into three groups of three consonants by longer temporal pauses between-groups than within-groups. We aim to clarify whether the same grouping effects found on forward serial recall generalize to backward serial recall, and further discuss how the backward serial recall results could contribute to current memory theories.

	Forward Grouped	Forward Ungrouped	Backward Grouped	Backward Ungrouped
Mean recall $\leq 2\%$	0	2	1	6
Total	43	43	43	43

Table 4.1: The number of participants rejected for analysis (Mean recall ≤ 2) versus total number of subjects in each condition. A chi-square test found differences between number of included subjects between Backward Grouped and Backward Ungrouped was significant (χ^2(df=1)=3.89, $p < 0.05$). No difference was found between Forward Grouped and Forward Ungrouped.

4.2 Methods

4.2.1 Participants

A total of 172 undergraduate students from introductory psychology courses at the University of Alberta participated in exchange for partial course credit. Participants gave informed consent, had normal or corrected-to-normal vision and learned English before age 6. Participants were run in groups of about 10–15, with random assignment to testing groups. Nine participants were excluded because they recall less than 2 items from the 9 item list. The number of excluded versus included participants in each condition is summarized in Table 4.1.

4.2.2 Materials & Procedure

To maintain continuity with previous studies (Chan et al., 2009; Liu et al., 2014), stimuli were 16 consonants (excluding S, W, X, and Z) from the English alphabet displayed in uppercase. Each list comprised 9 consonants drawn at random without replacement from the stimulus pool. Probability was equal for each consonant/serial-position combination. All participants were tested using a group of 15 computers (custom-built PCs) with identical hardware, identical Samsung SyncMaster B2440 monitors and Logitech K200 keyboards, to minimize hardware precision variability in our between-subject design (Plant & Turner, 2009).

The experiment was implemented with the Python Experiment-Programming Library (PyEPL; Geller et al., 2007) and modified from Chapter 2's (Liu et al., 2014) judgements of relative order experiment replacing the JOR test with a serial-recall test. Participants were

randomly assigned to one of four testing groups in a 2×2 design (Instruction ["forward", "backward"] by Grouping ["grouped", "ungrouped"]). Depending on the instruction, participants were either asked to type the list in forward or backward order: (a) Excerpt from "forward" instruction: "...you will be asked to type the list you just saw, starting from the first letter and ending with the last letter. In other words, type the list in forward order. ..." (b) Excerpt from "backward" instruction: "...you will be asked to type the list you just saw, starting from the most recent letter and end with the first letter. In other words, type the list in backward order. ...". Each trial began with a fixation asterisk, '*', in the center of the screen, followed by a consonant list presented sequentially in the center of the screen. Items were presented for 500 ms each. The "ungrouped" group had a constant inter-stimulus interval (ISI) of 350 ms, whereas "grouped" group had an ISI of 950 ms between items 3 and 4, and between items 6 and 7, to create a longer temporal gap between-groups, and an ISI of 150 ms for all other transitions. The ISIs were selected to maintain a constant total presentation time of 7300-ms between "grouped" and "ungrouped" group. After a 2500-ms delay, participants were cued with an input line and a text reminder to type the list either in forward or backward direction. Participants could not backtrack to edit entered consonants, and they terminated response by pressing the ENTER key. All letters of the English alphabet were accepted as input, and all typed letters stay on the computer screen until ENTER was pressed. After a 500-ms delay, participants could press a key to start the next trial.

Data analysis

We analyzed our data with linear mixed effects (LME) models (Baayen et al., 2008; Bates, 2005). We adopted LME analysis because compared to ANOVA, LME can fit individual responses without the need for averaging of the data, and protects against Type II error due to increased power (Baayen et al., 2008; Baayen & Milin, 2010). LME analyses were conducted in R (Bates, 2005), using the LME4 (Bates & Sarkar, 2007), LanguageR (Baayen, 2007) and LMERConvenienceFunctions (Tremblay, 2013) libraries. The "lmer" function was used to fit the LME model. The "Anova" function from Companion to Applied Regression package (Fox & Weisberg, 2011) was used to conduct Wald chi-square test for the best fitting models. The "mcposthoc.fnc" function was used to conduct posthoc analysis.

We used both serial position and output position for accuracy analysis. The latency measure was the time difference for recalling two subsequent items, thus we analyzed latency

data based on output position. Recall latency was excluded when an one of the two items was an error.

Three factors were used as fixed effects to predict log-transformed latency: Direction, Grouping and Output Position (serial position based on recall order). Subject was included as a random effect on intercept. Grouping and Output Position were treated as categorical factors. Recall latency was log-transformed to reduce skewness.

The accuracy data were fitted with logistic regression as it is a binary variable ("correct" vs. "incorrect"). Accuracy was analyzed based on Serial Position (i.e., based on study order); thus we used Grouping, Direction and Serial Position as fixed effects predicting accuracy.

LME estimated random effects first, followed by fixed effects. In the results tables, the "Estimate" column reported the corresponding regression coefficients, along with their standard errors. For the purposes of reporting the LME results, "grouped" condition, "forward" direction, Input/Output Position 1, were set as the reference levels for the Grouping and Input/Output Position. Input position is identical to serial position, which we will use both terms interchangeably.

The best fits of LME models were obtained by conducting a series of iterative tests comparing progressively simpler models with more complex models using the Bayesian Information Criterion (BIC), as was done by Chapter 2 (Liu et al., 2014), using LMERConvenienceFunctions (Tremblay, 2013).

4.2.3 Results

The number of excluded participants was higher in Backward Ungrouped compared with Backward Grouped (Table 4.1, indicating grouping reduced task difficulty for backward recall. Caution should be made when interpreting results from the included participants in ungrouped backward recall, as the high performers may have already started grouping the list spontaneously. For the included participants, we first inspected the effect of Grouping and Direction collapsing across Input/Output positions. The best fitting LME model for accuracy was reported in Table 4.2 (left panel), showing a main effect of Grouping and Direction, with no significant Grouping $\times$ Direction interaction (see Figure 4.1a). The "grouped" was more accurate than "ungrouped" and the "forward" direction was more accurate than the "backward" direction. The best fitting LME model for latency was reported in Table 4.2 (right panel) showing a main effect of Grouping and Direction, with no significant Grouping

	Accuracy	Latency
Main effects	Estimate (SE)	
Intercept	0.104 (0.088)*	6.17 (0.048)*
Grouping	-0.36 (0.10)*	0.18 (0.067)*
Direction	-0.31 (0.10)*	0.079 (0.095)*

Table 4.2: The best-fitting LME model for accuracy (left panel) and latency (right panel), collapsing across input/output positions. The "Estimate" column reports the corresponding regression coefficient, along with its SE (standard error). Significant effects are denoted * - $p < 0.05$.

× Direction interaction (see Figure 4.1a). The main effects for latency suggested no speed-accuracy tradeoffs (see Figure 4.1b). "Grouped" group was faster than "ungrouped" group and "forward" direction latency was faster than "backward" direction. In sum, grouping enhanced serial recall on both memory accuracy and access speed, and forward recall was performed better than backward recall.

Effects of grouping

Because grouping and Direction did not interact, we investigated the input/output position effect separately collapsing through each factor. We first looked at the effects of Grouping on forward and backward serial recall, then looked at the effects of Direction on "grouped" and "ungrouped" lists.

First we looked at results from forward serial recall. In line with previous findings, "ungrouped" group showed a smooth accuracy serial position curve, with a strong primacy effect and a weak recency effect (figure 4.2a). Compared with the "ungrouped" group, the "grouped" group showed non-smooth accuracy transitions at group boundaries, and an enhanced accuracy in the last group. We quantified the effects of Grouping with posthoc comparisons of "grouped" versus "ungrouped" lists at individual serial positions. The best fitting LME model (Table 4.4a) found main effect of Grouping ($p < 0.05$), a significant effect of Serial Position ($p < 0.05$) and a significant interaction of Grouping × Serial Position ($p < 0.05$). Confirming our qualitative observations, posthoc comparisons found grouping enhanced accuracy at Serial Positions 3, 6, 7, 8 and 9 while not affecting accuracy at Serial Positions 1, 2, 4 and 5.

We also replicated the classic effects of grouping on error types, where grouping reduced

a **b**

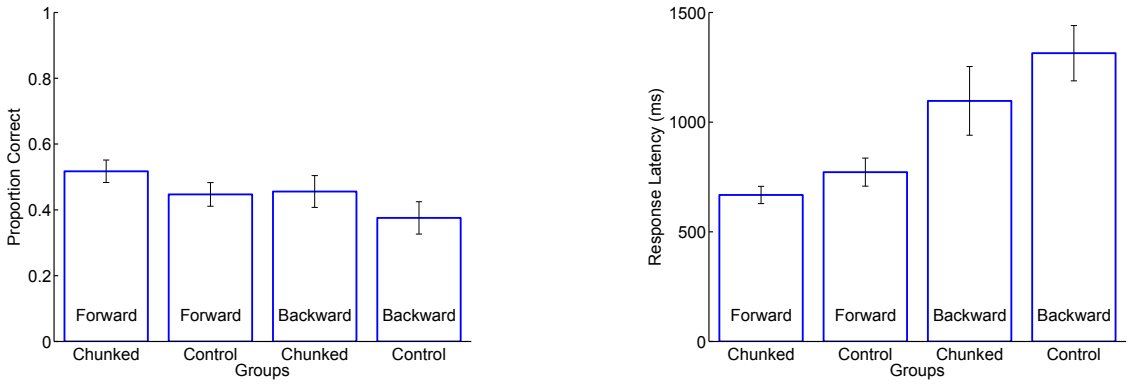


Figure 4.1: Mean latency (panel a) and accuracy (panel b) as a function of group. The error bars are 95% confidence intervals.

	Forward		Backward	
	Grouped	Ungrouped	Grouped	Ungrouped
Adjacent Transposition	32.11%	39.99%	12.06%	16.21%
Interposition	25.75%	14.32%	6.25%	6.19%

Table 4.3: Proportion of adjacent errors and interposition errors for each condition. Cell values are calculated by dividing the number of adjacent transposition and interposition errors by total number of errors per condition

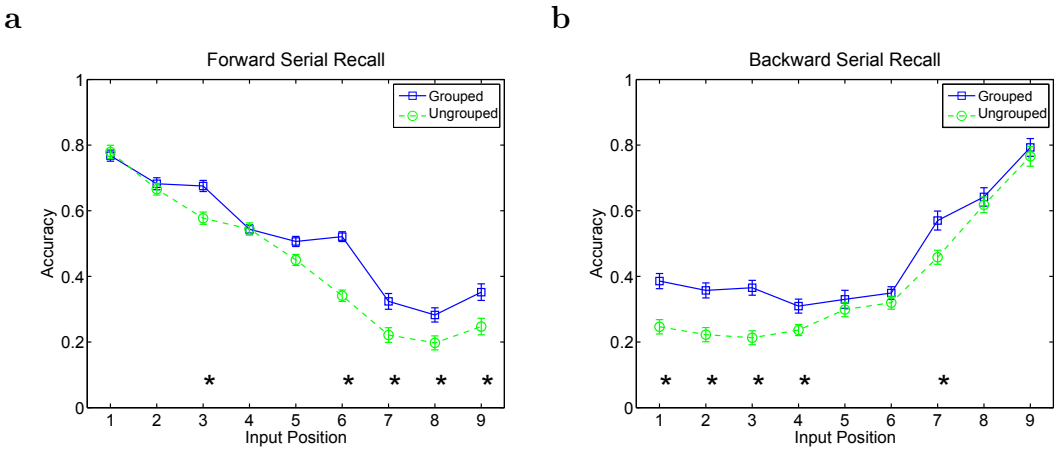


Figure 4.2: Recall accuracy as a function of serial position for forward recall (panel a) and backward recall (panel b). Significant difference between "grouped" and "ungrouped" group is denoted by "*" ($p < 0.05$)

adjacent transposition errors and increased interposition errors for forward serial recall (Table 4.3)

The latency output position curve for both "grouped" and "ungrouped" group showed an initial slow response to initiate the recall (Figure 4.3a). The "grouped" group had longer recall latencies cross group boundaries than the "ungrouped" group. The analysis of the best fitting LME model (Table 4.5a) found a main effect of Grouping ($p < 0.05$) with "grouped" faster than "ungrouped", a significant effect of Ouput Position ($p < 0.05$) and a significant interaction of Grouping $\times$ Output Position interaction ($p < 0.05$). Posthoc comparisons found grouping only increases latency at output position 4 and 7, where the latency is across boundary between two adjacent groups, confirming the visual impression.

The accuracy serial position curve, error patterns and latency output position curves are in line with predictions based on hierarchical position codes. Chaining models, without further assumptions, may produce mirror images of serial position curves, but cannot yet

		χ^2	df
a			
Forward Recall			
	Grouping	7.58 *	1
	Serial Position	6497.19 *	8
	Grouping × Serial Position	219.42 *	8
b			
Backward Recall			
	Grouping	5.55 *	1
	Serial Position	6470.10 *	8
	Grouping × Serial Position	216.18 *	8

Table 4.4: The best-fitting LME model for Forward (panel a) and Backward (panel b) recall accuracy (proportion correct). The χ^2 column reports χ^2 from Wald test. The df column reports corresponding degrees of freedom. Significant effects are denoted * - $p < 0.05$.

		χ^2	df
a			
Forward Recall			
	Grouping	11.17 *	1
	Output Position	16124.96 *	8
	Grouping × Output Position	1144.98 *	8
b			
Backward Recall			
	Grouping	8.15 *	1
	Output Position	12222.56 *	8
	Grouping × Output Position	1063.94 *	8

Table 4.5: The best-fitting LME model for Forward (panel a) and Backward (panel b) latency (ms). The χ^2 column reports χ^2 from Wald test. The df column reports corresponding degrees of freedom. Significant effects are denoted * - $p < 0.05$.

account for the increase of interposition errors for "grouped" lists.

Backward serial recall

We now turn to the analysis of backward serial recall data. Comparing to the "ungrouped" group, the "grouped" group showed non-smooth accuracy transitions at group boundaries (serial positions 3 and 6), and enhanced accuracy in the last group (serial positions 7, 8 and 9). We quantify the effects of grouping by posthoc comparisons of "grouped" versus "ungrouped" subjects at individual serial positions. The best fitting LME model (Table 4.4a) found a main effect of grouping ($p < 0.05$), a significant effect of Serial Position ($p < 0.05$) and a significant interaction of Grouping × Serial Position ($p < 0.05$). Confirming our

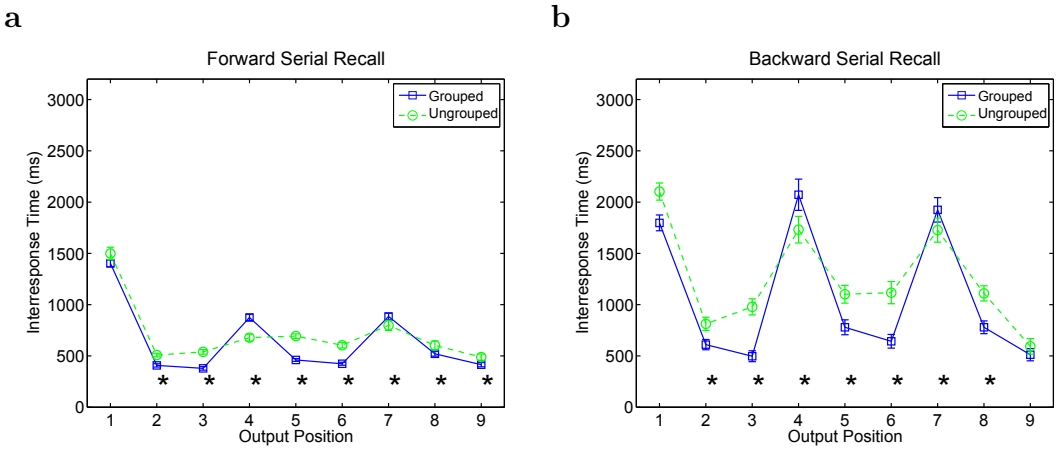


Figure 4.3: Recall latency as a function of output position for forward recall (panel a) and backward recall (panel b). Significant difference between "grouped" and "ungrouped" group is denoted by "*" ($p < 0.05$) and a non-significant trend is denoted by '.' ($p < 0.10$)

		χ^2	df
a			
Grouped			
	Grouping	3.65 *	1
	Output Position	545.42 *	8
	Grouping × Output Position	5049.53 *	8
b			
Ungrouped			
	Grouping	4.06 *	1
	Output Position	623.61 *	8
	Grouping × Output Position	7180.3 *	8

Table 4.6: The best-fitting LME model for "grouped" (panel a) and "ungrouped" (panel b) recall accuracy (proportion correct). The χ^2 column reports χ^2 from Wald test. The df column reports corresponding degrees of freedom. Significant effects are denoted * - $p < 0.05$.

		χ^2	df
a			
Grouped			
	Grouping	5.67 *	1
	Output Position	202220 *	8
	Grouping × Output Position	519.68 *	8
b			
Ungrouped			
	Grouping	17.08 *	1
	Output Position	9088.14 *	8
	Grouping × Output Position	407.78*	8

Table 4.7: The best-fitting LME model for "grouped" (panel a) and "ungrouped" (panel b) latency (ms). The χ^2 column reports χ^2 from Wald test. The df column reports corresponding degrees of freedom. Significant effects are denoted * - $p < 0.05$.

qualitative observations, posthoc comparisons found grouping enhanced accuracy at Serial Positions 3, 6, 7, 8 and 9 while not affecting accuracy at Serial Positions 1, 2, 4 and 5.

The backward serial recall serial position curves resembled mirror images of the forward serial recall serial position curves. The "ungrouped" group showed a smooth accuracy serial position curve with a strong recency effect and a weak primacy effect (Figure 4.2b). Compared to the "ungrouped" group, the "grouped" group showed non-smooth accuracy transitions at group boundaries, and enhanced accuracy at the first group. The best fitting LME model (Table 4.4b) found a main effect of Grouping ($p < 0.05$), a significant effect of Serial Position ($p < 0.05$) and a significant interaction of Grouping × Serial Position ($p < 0.05$). Posthoc comparisons found grouping enhanced accuracy at Serial Positions 1, 2, 3, 4, 7 while not affecting accuracy at Serial Positions 5, 6, 8, and 9.

The latency pattern for both "grouped" and "ungrouped" backward serial recall showed "scalloped" serial position curves, with slower performance at output positions 1, 4 and 7. The "grouped" group has slower latencies and showed a more pronounced "scalloped" effect than "ungrouped" (Figure 4.3b). The similar serial position curve between "grouped" and "ungrouped" backward serial recall lists suggests for backward serial recall, participants spontaneously grouped the list into 3-item groups, consistent with evidence of spontaneous grouping (e.g., Martin & Noreen, 1974; Madigan, 1980; Jou, 2011). The best fitting LME model (Table 4.5b) found a main effect of Grouping ($p < 0.05$), a significant effect of Output Position ($p < 0.05$) and a significant interaction of Grouping × Output Position ($p < 0.05$).

Posthoc comparisons found "grouped" group had faster latencies at Output Position 2, 3, 5, 6, 8, slower latencies at Output Position 4, 7 and not affecting latency at Output Position 1 and 9.

For the error patterns, we found the adjacent transposition errors were higher for the ungrouped list than the grouped list, in line with the forward serial recall results (Figure 4.3). However, we did not find an increase of interposition error for the Backward Grouped list. The lack of increased interposition errors may have been due to spontaneous grouping in the ungrouped backward serial recall, indicated by the scalloped response time output position curve (Figure 4.5a).

Effects of recall direction

As suggested by the best fitting overall LME model, forward recall was faster and more accurate than backward recall. To further understand the effects of recall direction, we looked at recall direction separately for "ungrouped" lists and "grouped" lists. When serial position was used to collapse accuracy responses, forward recall produced a bigger primacy effect and a smaller recency effect for both "ungrouped" (Figure 4.4a) and "grouped" lists (Figure 4.4b). The serial position curve for backward serial recall was almost a mirror image or forward serial recall, suggesting the serial position curve was dominated by output order instead of learning order. When plotted in output order (Figure 4.4c and "grouped" lists (Figure 4.4d), both "ungrouped" and "grouped" groups found forward serial recall had higher accuracy than backward serial recall only at the middle of the list (output positions 3, 4, 5 and 6).

The latency output position curve for "ungrouped" list showed forward recall had an overall faster latency, and was smoother than the backward recall output position curve (Figure 4.5a). The scalloping in the backward recall output position curve suggested participants spontaneously group the lists into 3 groups of 3 items. When looking at "grouped" group, where group of 3-items were induced, forward and backward serial recall were qualitatively similar and showed scalloping (Figure 4.5b), where the first item in a group (position 1, 4 and 7) was slowest compared with other items within the same group. The pause before recalling each group was longer for backward serial recall than forward serial recall, where latency for most other items was at the same level. Looking at both accuracy and latency results, the backward-recall latency showed a clear pattern of grouping as can be seen in

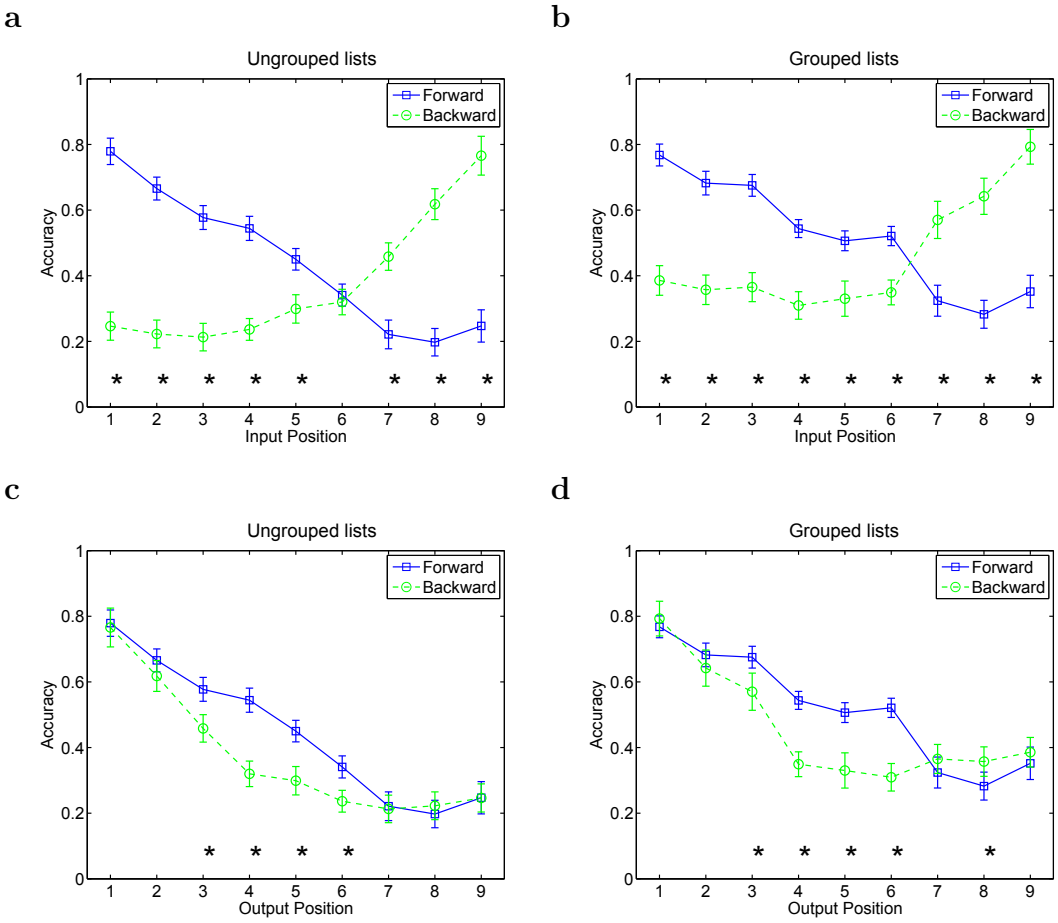


Figure 4.4: Recall accuracy as a function of serial position for "ungrouped" (panel a) and "grouped" list (panel b) and as a function of output position for "ungrouped" (panel c) and "grouped" list (panel d) Significant difference between recall direction is denoted by "*" ($p < 0.05$) and a non-significant trend is denoted by '.' ($p < 0.10$)

extra pauses before recalling first item of each group; however, the grouping strategy did not seem to help backward serial recall as much as forward serial recall, especially in the middle group. Both latency and accuracy patterns suggested when recalling in the backward direction, participants might have more difficulty accessing contents of each group, particularly in the middle of the list, causing slower latency and lower accuracy.

4.2.4 Discussion

In this experiment, we found temporally induced grouping on backward serial recall is quantitatively very similar to effects of grouping on forward serial recall when aligned using the

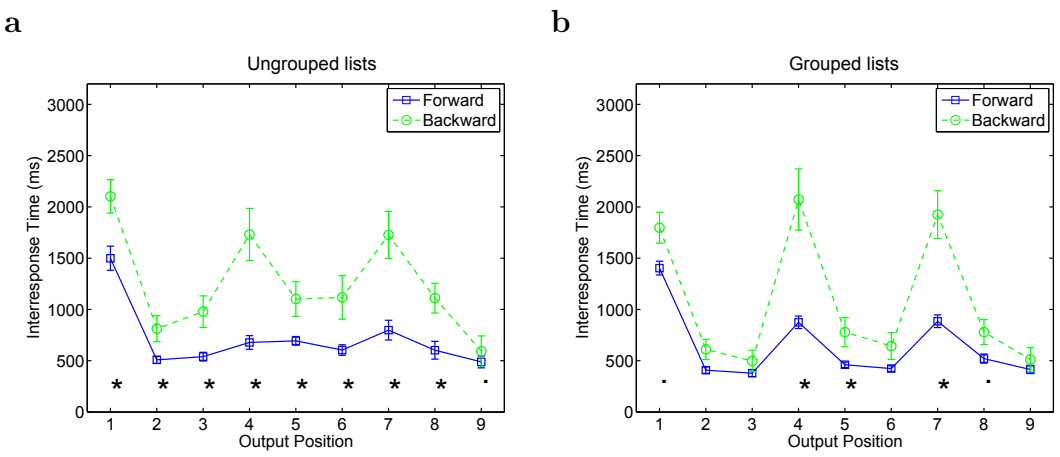


Figure 4.5: Recall latency as a function of output position for "ungrouped" (panel a) and "grouped" list (panel b). Significant difference between recall direction is denoted by "*" ($p < 0.05$) and a non-significant trend is denoted by '.' ($p < 0.10$)

output order, despite the main effect that forward serial recall is overall more accurate and faster than backward recall. The general effects of grouping across recall directions share some key features: a) increased transposition errors, b) recall latency for the first outputted item of each group is slower, accompanied by a slower latency than ungrouped control, c) recall accuracy for the last recalled group is higher than ungrouped control and d) latency decreases monotonically as within-group output position increases. We first discuss the compatibility of our results with our major hypothesis, then further discuss its implications on different classes of memory models.

Grouping effects on backward serial recall

With direct comparison between "grouped" and "ungrouped" groups for both recall directions, and detailed analysis of latency patterns, we showed that for backward serial recall, grouping induces the same effects on error patterns, and induces the same qualitative "scalloped" patterns as found on forward serial recall, for both the accuracy and latency. The latency pattern suggests a monotonic decrease in latency as within-group position increase. This pattern is consistent with the hypothesis that group position can be directly read out in backward order, and against the idea that group-level position codes were first retrieved and then reversed, before the reversed list is outputted in the reversed sequence. The latency pattern is, however, also compatible with the forward scan-and-drop strategy (Conrad,

1965), as latency for first item in each group is slower in backward recall than forward recall and latency drops as within-group position increase. However, the accuracy pattern, when recall accuracy is aligned by output order, does not suggest the backward serial recall is conducted with a scan-and-drop strategy, as backward recall does not show an enhanced primacy effect versus forward serial recall (Thomas et al., 2003; Haberlandt et al., 2005).

Our results are consistent with the hypothesis that positions or chains could be encoded or retrieved in reverse order; however, we could not rule out whether participants are encoding the list positions in backward order, because recall direction is pre-cued. We argue it is unlikely backward serial recall participants encode the list in backward order, as previous results (e.g., Bireta et al., 2010) found pre-cued and post-cued recall does not affect behavioural patterns of forward and backward serial recall in the absence of grouping manipulation.

Effects of Recall direction

The performance differences in forward and backward serial recall have led to different theoretical explanations. In one view, backward recall is thought to be more complex than forward serial recall (for a review on this topic see St. Clair-Thompson & Allen, 2013; Rosen & Engle, 1997). The hypothesis is that backward recall generally requires the same resources as forward serial recall, but also requires an additional attentional effort to mentally transform the sequence. There is considerable support that backward recall is more task demanding than forward recall, that **a**) backward and forward serial recall have been found to load on different short-term memory factors in factor analysis (e.g., Alloway, Gathercole, & Pickering, 2006), **b**) backward recall is more sensitive to effect of aging and brain dysfunctions relevant to executive control (Reynolds, 1997), and **c**) backward recall exhibits greater bilateral activation in the dorsal lateral prefrontal cortex than forward serial recall (Gerton, Brown, Meyer-Lindenberg, Kohn, & Holt, 2004). We found backward recall for both "grouped" and "ungrouped" lists was slower than forward recall, and backward recall for both "grouped" and "ungrouped" lists was less accurate than forward recall at the mid-list output positions.

Another view comes from evidence suggesting backward and forward recall may use different underlying representations, that backward recall utilizes a visual spatial code whereas forward recall utilizes a phonological code. Li and Lewandowsky (1995) found backward recall, not forward recall, was disrupted by presenting items in different spatial location

and enhanced by intralist visual similarity. If backward recall is not based on phonological representations, benchmark effects based on phonological representations should not be present for backward recall. In line with this, some studies failed to find benchmark effects such as phonological similarity, word length, articulatory suppression and irrelevant speech in backward serial recall (Baker et al., 2012; Bireta et al., 2010; Surprenant et al., 2011). Ritchie et al. (2015) conducted a meta-analysis of the phonological benchmark effects and found those effects were severely attenuated for the first two items outputted from backward recall, suggesting at least for the immediately outputted items, backward recall may not rely on phonological representations. On the other hand, backward recall may be more similar to forward recall than different. Guérard et al. (2012) found the same benchmark effect of phonological similarity, word length, articulatory suppression and irrelevant speech reported to be absent in backward serial recall (Baker et al., 2012; Bireta et al., 2010; Surprenant et al., 2011) using a much larger word pool size. Ritchie et al. (2015) also found the phonological benchmark effects re-emerges for list items outputted after the first two items, suggesting a large part of backward serial recall is similar to forward serial recall. In addition, researchers found recall direction has no effect on predicting standardized test scores (Rosen & Engle, 1997).

Our data generally agree with the hypothesis that backward recall is more task demanding than forward recall. However, we suggest the effects of grouping are also diagnostic of whether participants could change the way they encode the list if they are preparing for backward recall, versus forward recall. It is unlikely the effects of grouping are identical if the underlying representation differ between forward and backward serial recall. We found that grouping did not interact with recall direction. This is similar to the results J. A. Anderson et al. (1998) obtained. Perhaps their spatial cueing, which we mentioned in the introduction, was not the critical feature of the methods that produced similar effects of grouping on backward as on forward recall. The similarity between grouped lists across recall direction challenges Li and Lewandowsky's (1995) interpretation and adds evidence favouring the view that forward and backward recall share similar underlying representations.

4.2.5 Models and theories

Implication for positional coding

The effects of grouping were found to be similar between forward and backward recall when data was aligned using output position. The symmetrical grouping effect across recall direction challenges the idea that each group has to be coded in forward direction and additional processing may be required for backward recall. Our results are compatible with the idea that each group can be retrieved by direct readout of the to-be-recalled items, either in forward or backward direction. This direct readout mechanism should be relatively easy to adapt to positional coding models, as items are directly associated with position codes, and direct readout direction could be manipulated by explicit rules of cueing with position codes in reverse order. Our latency results are compatible with the assumption that additional processing is required to access contents within each group (Farrell, 2012b).

Models assuming a hierarchical representation generally agree that there is a position code at the level of each group in addition to a list-level position code (Farrell, 2012b). The two-level representation of position codes provides more precision of the to be recalled items, thus increases recall accuracy. Although we found a main effect of recall accuracy, the enhancement of recall is only at the last outputted group and last outputted item of each group. This pattern is inconsistent with previous findings that grouping increases accuracy across all serial positions (Hitch et al., 1996), and is consistent with Parmentier et al.'s (2006) finding that manipulation of the pre- and post- item interval had little effect on order reproduction accuracy .

One possible explanation is that we equated the total presentation time for "grouped" and "ungrouped" lists, and the overall grouping advantage found in previous studies could be partially attributed to the longer overall presentation for "grouped" than "ungrouped" lists(Hitch et al., 1996). When the overall presentation time for "grouped" and "ungrouped" groups are equal, grouping predominantly enhances accuracy for serial positions at the middle of a list or a group, and produces little accuracy enhancements at the beginning and end of a list or a group (e.g., Maybery et al., 2002; Ng & Maybery, 2005; Farrell & Lelièvre, 2012). Another possible explanation is that the assumption that two-level positional codes enhance precision of memory does not always hold. It is reasonable to expect the two-level position codes competing for the same cognitive resources, and the availability of the

cognitive resources may become a bottle neck of memory enhancement. Thus, depending on the cognitive capacity of participants, the difficulty of the memory task, we may find mixed effects of the two-level position codes.

ACT-R

J. A. Anderson et al. (1998) expanded the ACT-R model to specifically account for the grouping effects. The ACT-R model assumes a list is represented as sets of groups and each group is represented as a set of items. There is a high-level code representing each group's position in the list, and the group size, and another lower-level code representing each item's position within each group, thus a two-level hierarchical code is used. A production rule is used for retrieval of items from the list, which retrieves a specific item in the list by first retrieving the information about all groups (group position code and group size), then unpacks the specific group enclosing the items. Thus, this model predicts slower latencies for the first item and first item within each group, because of the additional process of retrieving group-level information. However, this model predicts the items following the first item of each group should be recalled at a constant rate. J. A. Anderson et al.'s (1998) version of ACT-R is not consistent with our data, where latency speeds up following the first item of each group. Our results are consistent with Ng and Maybery's (2005) version of ACT-R, where they suggest in addition to the group-level knowledge units and item-level knowledge units to predict latency, the number of items left for recall, competition across groups, and additional time required to switch task from study to recall items, are all important factors to predict latency. The number of items left for recall would predict a steady decrease of latency after the first item in each group (e.g., Bireta et al., 2010), because of reduced response competition. The competition across group, a factor represent swapped group codes, is required to account for increased transposition errors, and additional time for task switching, a factor represent additional time to switch from forward recall to backward recall, is to account for the slow latency for the first recalled item (e.g., Maybery et al., 2002).

To account for the effects of grouping for backward recall, one might assume the production rule can be implemented in reverse order: First retrieve the last group (m), and retrieve the last item within the group (n) and followed by the $n-1th$ item, when all items in group m have been recalled, the process proceeds to group $m-1$, and this process could continue until all items are retrieved for each group. This model would predict the effects of grouping

are nearly identical between recall directions.

SIMPLE

SIMPLE (Brown et al., 2007) is a scale-invariant model that assumes discrimination within a multidimensional psychological space drives the memory recall. The primary dimension for serial order memory is assumed to be the temporal dimension, where presentation rate and retention interval determines the discriminability of list items. SIMPLE has not been applied to latency data. However, SIMPLE models the accuracy effects of grouping by assuming a within-group position dimension in addition to the temporal dimension. For example, the within-group dimension for a list with three groups of three items would be coded as 123123123, where the within-group dimension is combined with the temporal dimension to determine serial position curves. The within-group position dimension is usually assigned with more weights than the temporal dimension, to model the "scalloped" pattern. SIMPLE predicts items with a relatively larger temporal isolation should be recalled better. This would lead to the prediction that the first and last item of each group should be facilitated and the overall serial position pattern should show a primacy effect and as well as a recency effect. The inclusion of an additional dimension also predicts the overall accuracy for "grouped" group to be higher than "ungrouped", where accuracy at all serial positions should not be equal between groups. In addition, SIMPLE assumes forward and backward recall share the same encoding phase; however, SIMPLE still predicts a steeper serial position curve with more recency effect for the backward recall. This is because time proceeds for each item recalled and depending on where in the list to start recall, the study-test interval change affect discriminability of the other items (Li & Lewandowsky, 1993).

As mentioned in the positional coding models discussion, our finding that grouping does not enhance the recall of the first item of a group of the first two groups challenges SIMPLE and distinctiveness based accounts in general. Our results are in line with Parmentier et al.'s (2006) finding that manipulation of the pre- and post- item interval had little effect on accuracy. In addition, forward serial recall and backward serial recall have differential study to test interval per each recalled item (Li & Lewandowsky, 1993). During recall, the study and test interval is approximately constant for forward serial recall, whereas study and test interval increase for backward serial recall. SIMPLE would predict the overall recency pattern to be very different between recall direction, regardless whether we align the data

by input or output positions. We found the serial position curve for "grouped" lists are symmetrical, inconsistent with SIMPLE's prediction.

In sum, we suggest SIMPLE needs further assumptions to account for our results. Specifically, to explain the finding that grouping does not always enhance the first item of a group and grouping effects are very similar between instructions, when output order is used to align data.

TODAM

Different from the positional coding accounts and SIMPLE, TODAM relies on inter-item associations. TODAM (Lewandowsky & Murdock, 1989) assumes each retrieved item serves as the cue for the next item, and there is associative symmetry of the component pairs of the association chain. For example, given a pair A–B, recalling A given B should be equally accurate as recalling B given A (Asch & Ebenholtz, 1962; Kahana, 2002; Köhler, 1947). One way of implementing backward serial recall is to start recall from the last item of the list. This would predict the forward and backward recall pattern should be the same across recall direction, in line with our "grouped" group results.

However, our serial position effect challenges TODAM. If we assume grouping enhances within-group association and reduces between-group associations, it is predicted that within each group, the first item's recall accuracy is lowest followed by higher within-group recall accuracies for the within-group items, and because of associative symmetry, this would apply to both forward and backward recall. In our study, we found selective accuracy enhancement for the last item recalled in each group. Additional work need to be done to understand how can an associative chaining account enhanced accuracy for the last item of each group, but not the first item. We suggest this may be due to compound cueing within the group, that the third item is retrieved by a combination of an adjacent association from the second item and a remote association from the first item.

Conclusion

In conclusion, our results demonstrate the effect of grouping is very similar in forward and backward serial recall when latency and accuracy are aligned by output order. This suggests grouping influences retrieval processes, but may not influence the way order is encoded. Our results pose challenges to models that rely on temporal coding and temporal discriminability

(e.g., SIMPLE; Brown et al., 2007), and to models rely on the associative chaining mechanism (e.g., TODAM Murdock, 1995). Although we pose challenges to the positional coding models. A simple modification of positional coding models, such as the ACT-R model, might be able to accommodate our results, as we showed position code should be able to directly accessed in backward order in backward recall.

Chapter 5

Effects of Grouping on Judgements of Relative Order

5.1 Introduction

Memory of order is prevalent in our everyday activities. We are often required to process temporal order information such as recalling plots of a movie, determining who arrived late for a meeting. One way of testing temporal order memory is asking participants to judge relative temporal order between two items from a sequence, which we call judgements of relative order (JOR) (see Chapter 1). For example, given a pair A-B from a list ABCD, participants were asked to judge which item came earlier/later. JORs were extensively studied with a instruction asking "which item is more recent" (Hacker, 1980; Muter, 1979; Yntema & Trask, 1963). Chan et al. (2009) found on subspan lists of consonants (list length 3-6), the instruction wording "Which item came earlier?" versus "Which item came later?" showed a crossover interaction with the probe serial positions, characterized by selective facilitation of response time and error rate towards the beginning of the list by the "earlier" instruction, or towards the end of the list by the "later" instruction (Chan et al., 2009). Chapter 2 (Liu et al., 2014) replicated this interaction between instruction and serial position on supraspan lists (list length 10 words, list length 8 consonants), and Chapter 3 further showed the JOR congruity effect could be found on English alphabet (see also Jou & Aldridge, 1999). In this chapter, we ask whether the JOR congruity effect could be affected by grouping a long list into smaller lists, and whether JOR results are theoretically compatible with comparative judgements and serial recall.

In serial recall, participants are asked to recall a studied list in order. It has been

reported that participants spontaneously group the list into smaller lists (e.g., Madigan, 1980; Jou, 2011), and a grouping strategy can be easily induced though various methods such as a temporal pause (e.g., Hitch et al., 1996; Frankish, 1985; Maybery et al., 2002), spatial grouping (e.g., J. A. Anderson et al., 1998; Parmentier & Maybery, 2008), voice cues (Frankish, 1989; Parmentier & Maybery, 2008), and an overt instruction to form groups (e.g., Farrell, 2012a). Grouping facilitates overall recall accuracy, and induces within-group mini serial position curves that each show primacy and recency effects (Henson, 1998; Lee & Estes, 1981; Maybery et al., 2002; Ng & Maybery, 2002, 2005; Wickelgren, 1967; Ryan, 1969a). Grouping also affects the overall pattern of errors (Ryan, 1969a, 1969b). Comparing "grouped" with "ungrouped" list, grouping reduces adjacent transposition errors, that is when an item is recalled at an adjacent serial position. In exchange, grouping increases interposition errors, that is when an item is recalled in the wrong group but with the correct within-group position (Farrell & Lewandowsky, 2004). For example, for a list of 9 items comprised of three groups: 3-3-3, item 2 being recalled in position 3 would be an adjacent transposition error, whereas item two being recalled in position 5 would be an interposition error. Grouping also produces extended pauses before outputting each group, creating a longer inter-response time between-group (Chapter 4; Maybery et al., 2002; Ng & Maybery, 2002, 2005; Farrell & Lelièvre, 2012).

The effects of grouping on JORs have not been previously studied; however we can speculate on the grouping effects based on a closely related paradigm, the comparative judgements, where two alternative stimuli are compared along an attribute dimension that is either directly perceived, or accessed from memory. Comparative judgements data show three benchmark effects: a) a serial position effect, where items at the beginning and end of the continuum are faster to respond; b) a distance effect, where a greater difference between two items along a underlying continuum predicts faster response time (Moyer & Bayer, 1976; Banks, 1977); c) a semantic congruity effect where response time is faster when the semantic attribute is congruent with the wording of the instruction and vice versa when the semantic attribute is incongruent with the wording of instruction (Banks, 1977). For example, when asked which animal is larger between rhino and elephant, the response time is faster than the same judgement between cat and rabbit, because the "larger" question is congruent with the larger size of the rhino and elephant. Data on JORs exhibit all three benchmark effects found in comparative judgements, and it has been shown that these benchmark effects also

extend to the error-rate measure (see Chapter 2 (Liu et al., 2014) and Chapter 3). The similar behavioural findings between JORs and comparative judgements suggest the JOR task could be conceptualized as a comparative judgement task along the continuum of time (Jou, 2011; Brown et al., 2007).

In the comparative judgement literature, there is mixed evidence on how organizing the continuum in discrete groups would affect the benchmark effects. Some researchers (Holyoak & Patterson, 1981; Kosslyn et al., 1977) have suggested that the group label and the magnitude continuum could both be used for comparison, where the group labels use discrete codes and magnitude along a continuum uses analogue codes. For instance, for a list containing both animals and buildings, the group label ANIMAL and BUILDING could be used to make a relative size judgement between cat and house without having to compare the size of a particular member of ANIMAL and BUILDING (namely, cat and house). Alternatively, when the group labels are not used for judgements, the analogue sizes of cat and house will be used for comparison. The distance effect has been suggested to be a by-product of magnitude comparison; therefore when no magnitude comparison is required, there should be no distance effect (e.g., Holyoak & Patterson, 1981; Kosslyn et al., 1977). This theory leads to the prediction that within-group judgements should show a distance effect, whereas between-group judgements should show no distance effect, and between-group judgements should be faster or as fast as within-group judgements. Some studies showed a violation of this hypotheses that a distance effect could be found in between-group judgements (Howard, 1980; Kosslyn et al., 1977; Maki, 1982; Woocher et al., 1978) and between-group judgements could be longer than within-group judgements (Kosslyn et al., 1977; Maki, 1981, 1982; Woocher et al., 1978). However, the distance effect could be attenuated (Maki, 1981) or disappear (Kosslyn et al., 1977; Pohl, 1990; Pliske & Smith, 1979) when the groups are over-learned (Kosslyn et al., 1977), or from pre-existing semantic categories (Howard, 1980; Maki, 1981; Pliske & Smith, 1979; Sailor & Shoben, 1993; Shoben & Wilson, 1998), or when the serial position effect could be controlled (Pohl, 1990). This set of results suggests the group labels could be used by participants, but it is a less efficient strategy than performing relative magnitude comparison using the analogue codes, especially when groups are not already well learned and easily retrievable. Cech and Shoben (2001) suggest both strategies could compete with each other to maximize efficiency. Unlike the distance effect, grouping has not been found to influence the serial position effect in comparative judgements.

Comparative judgements produce an inverted-U shaped serial position effect, showing little difference between grouped and ungrouped lists (Woocher et al., 1978; Jou, 2005, 2011).

Our main goal is to identify the form of interaction between grouping and the congruity effect. To our knowledge, the comparative judgement literature has not tested how grouping could interact with the third benchmark effect – the congruity effect. We wonder how the congruity effect could be affected by grouping, as conflicting predictions could be made based on previous findings. If the congruity effect is like the distance effect, specific to magnitude comparisons and the group labels are discrete codes, one would not expect to find the congruity effect between groups. This would predict the congruity effect to be different between grouped and ungrouped lists. Because grouping has not been found to change the shape of serial position curve in comparative judgements, and the congruity effect is characterized by differential serial position curve slopes, one would expect the congruity effect is the same between grouped and ungrouped lists. In addition, we considered the possibility that the "earlier" instruction relies on covert forward serial recall and the "later" instruction relies on covert backward serial recall. In Chapter 4 we found the output position dominates the recall, which suggests the effect of recall direction is at the retrieval stage instead of the study stage. This hypothesis has been further supported by the finding that forward- and backward- serial recall serial-position-curves are close to mirror images, regardless of the grouping manipulation. Because grouping does not affect the interaction between forward and backward serial positions in serial recall, we expect grouping will not interact with the JOR congruity effect.

Our second goal is to relate the grouping literature on comparative judgement to grouping literature on serial recall, where grouping has a strong effect on serial position curves. The serial recall results support a hierarchical organization of memory for order, where positions are represented as relative to the list as a whole, or relative to the group (e.g., J. R. Anderson & Matessa, 1997; Farrell, 2012b; Henson, 1998; Lee & Estes, 1981). The hierarchical organization of memory for order led to the prediction that grouped lists have slower response times and higher accuracy than the ungrouped lists, regardless of the order memory task. This prediction is based on the assumption that processing an additional memory dimension takes additional effort, but the added dimension improves precision because information is available from both the group-level position codes and list-level position codes. However, unlike the serial recall results, the comparative judgement task shows smooth inverted-U

shaped curves regardless of experimental manipulations to induce grouping. Jou (2005) asked participants to learn signed height ranks from -7 to 7 associated with 15 people's names, testing whether the signed rank codes could induce grouping, as Jou (2005) considered the sign function like a group. Jou (2005) also used the cued recall task, where participants were given the position of the name and asked to recall the name associated with the position, and the absolute position identification task, where participants were given the names and asked to identify whether the name was the correct match for a given rank position. However, although both the cued recall and absolute position identification task showed non-smooth serial position curves that suggested participants spontaneously group the list into two categories, the comparative judgement task showed smooth U-shaped serial position curves that suggested no grouping. Consistent with the earlier research, Jou (2011) showed that a 16-item list of names associated with height ranks showed evidence of spontaneous grouping, only with the position-item cued recall, item-position cued recall, and absolute position identification task, but not with a relative order comparison task. The spontaneous grouping in the cued recall and the absolute position identification task may have been caused by the use of local reference points, such as the boundary items of a group. Jou (2011) argues that the comparative judgement task generates a smooth serial position curve spontaneously because the use of group-boundary information does not serve a useful function in comparative judgements as it does in cued recall and absolute position identification. For a list ABC DEF, item D could serve as a local reference because it is starting item of the second group. However, if D is used as a reference for comparing two items' order, the pair AB would be both earlier than D, the pair EF would be both later than D. Unless the probe items crosses D, knowing D's position would not benefit the judgement, facilitate recall or absolute serial position judgement by enhancing the resolution of serial positions near the boundary item. However, we suggest the use of group information could potentially benefit comparative judgements, with the use of a hierarchical positional code. Group labels could be considered as superordinate position codes, and comparisons could be performed by comparing the group labels. A response could be made based only on group labels if they are different. When group labels of probe items are identical, the response is based on comparison of subordinate position codes. A novel prediction following from the two-level hierarchical code could be made that when the group-level and list-level position code makes the same prediction (consistent), judgement accuracy should be facilitated, and

when the group-level and list-level position code makes opposite prediction (inconsistent), accuracy should be reduced due to interference, relative to the ungrouped behaviour. For example, for a list ABC DEF, the group-level position codes are 123 123, and the list-level position codes are 123456 (see Figure 1.2). An example of a "consistent" probe would be the probe AE, where the group-level position code is 1 and 2 for A and E, respectively, and the list-level position code is 1 and 5 respectively. Regardless of which code is used for the judgement, A has a code smaller than E. Following the same logic, B and D is a example of the inconsistent condition, where B has a smaller list-level position code than D, but B has a larger group-level position code than D. A and D is a example of the neutral condition, where both A and D share the same group-level position code, and no prediction can be made based on the group-level position code. It is possible that for the neutral condition, participants use group number or list-level position code to make judgement. We expect this prediction consistency effect to be present in both grouped and ungrouped lists, as evidence suggests that participants spontaneously subgroup lists, even when not instructed to do so (e.g., Madigan, 1980). In sum, we want to test whether theories developed from comparative judgements and theories developed from order memory research could both explain grouping effects and their interaction with the congruity effect.

Finally, we note two factors to consider when relating comparative judgements to serial recall. First the grouping experiment in comparative judgements are predominantly done with only two groups, with the exception of Pohl (1990). With only two groups, the distance effect and serial position effects are difficult to disentangle at the group level. The two-group experiments also provide little information on whether the group labels are discrete (Holyoak & Patterson, 1981) or also contain magnitude information. Positional coding models of the serial recall task suggest that for multiple groups, the group label is not discrete, but coded serially, and has been supported by a group code distance effect in Chapter 4, where we included three groups. A second factor is that the benchmark "scalloped" effects modelled by hierarchical position codes are based on the accuracy results. In a typical comparative judgement task, including Jou (2005, 2011), accuracy is trained to a ceiling criterion and only response time data could be used to infer grouping effects. However, the accuracy measure could be as diagnostic of the grouping effects as the response time pattern, if the list is not trained to an accuracy criterion. If the grouping effect is to be compared between paradigms, the accuracy data could provide insight in the underlying similarities and differences.

	Earlier Grouped	Earlier Ungrouped	Later Grouped	Later Ungrouped
	0	0	3	4
Total	36	36	36	35

Table 5.1: The number of participants rejected for analysis versus total number of subjects in each condition. A chi-square test found no differences between the "grouped" and "ungrouped" groups.

In this study, we use lists of 9 consonants with 3×3 item groups, induced by temporal pauses during list presentation. The identical presentation manipulation has been shown effective to induce grouping in forward and backward serial recall in Chapter 4. The only difference deviating from the serial recall experiment in Chapter 4 is that a JOR task instead of serial recall follows the list presentation. We use a 2×2 between-subject design (Instruction ["earlier"/"later"] $\times$ Group ["grouped"/"ungrouped"]). We present our results focusing on: a) whether the congruity effect generalizes across grouped lists, and b) whether grouping effects found in JORs can bridge theories generated from both comparative judgements and serial recall.

5.2 Methods

Participants

A total of 142 undergraduate students from introductory psychology courses at the University of Alberta participated in exchange for partial course credit. Participants gave informed consent, had normal or corrected-to-normal vision and learned English before age 6. Participants were run in groups of about 10–15, with random assignment to each testing group. Seven participants were excluded because more than 10 trials were faster than 200 ms. The number of excluded versus included participants in each condition is summarized in Table 5.1.

5.2.1 Materials & Procedure

Materials were identical to those used by Chapter 4. Stimuli were 16 consonants (excluding S, W, X, and Z) from the English alphabet displayed in capital letters. Each list comprised 9 consonants drawn at random without replacement from the stimulus pool. Probability was equal for each consonant/serial-position combination. All participants were tested using

a group of 15 computers (custom-built PCs) with identical hardware, identical Samsung SyncMaster B2440 monitors and Logitech K200 keyboards, to minimize hardware precision variability in our between-subjects design (Plant & Turner, 2009).

The experiment was created and run using the Python Experiment-Programming Library (Geller et al., 2007) and modified from the task used in Chapter 4. Participants were randomly assigned to one of four testing groups in a 2 × 2 design (Instruction ["forward", "backward"] by Grouping ["grouped", "control"]). Each trial began with a fixation asterisk, '*', in the center of the screen, followed by a consonant list presented sequentially in the center of the screen. Items were presented for 500 ms each. The "ungrouped" group had a constant inter-stimulus interval (ISI) of 350 ms, whereas the "grouped" group had an ISI of 950 ms among item 3, 4 and 6, 7 to create a temporal gap, and an ISI of 150 ms for all other list items. The ISIs were selected to maintain a constant total presentation time between the "grouped" and the "ungrouped" group. Instead of a serial recall task, we used a JOR task identical to that used in Chapter 2, Experiment 2.

Following each list presentation and a 2500-ms delay, participants were presented with a single probe consisting of two consonants from the just-presented list and were asked which item was presented earlier or later, depending on group, by pressing the ',' key (for the left-hand probe item) or the '/' key (for the right-hand probe item). Probes were pairs of items drawn from the just-presented list, and all possible combinations were equally probable and counterbalanced within subject. Each response was followed by a 500-ms delay before participants could press any key to start the next trial. Participants received slightly different instructions for each group: (a) Excerpt from "earlier" instruction: "...judge which of the two consonants came earlier on the list you just studied. Press the '/' key if the earlier item is presented on the right side of the screen and the '.' key if the earlier item is on the left side of the screen. ..." (b) Excerpt from "later" instruction: "...judge which of the two consonants came later on the list you just studied. Press the '/' key if the later item is presented on the right side of the screen and the '.' key if the later item is on the left side of the screen...". Participants were instructed to respond as quickly as they could without compromising accuracy.

5.2.2 Data analysis

We analyzed our data with linear mixed effects (LME) models (Baayen et al., 2008; Bates, 2005). We adopted LME analysis because compared to ANOVA, LME can fit individual responses without the need for averaging of the data, and protects against Type II error due to increased power (Baayen et al., 2008; Baayen & Milin, 2010). LME analyses were conducted in R (Bates, 2005), using the LME4 (Bates & Sarkar, 2007), LanguageR (Baayen, 2007) and LMERConvenienceFunctions (Tremblay, 2013) libraries. The "lmer" function was used to fit the LME model. The "Anova" function from Companion to Applied Regression package (Fox & Weisberg, 2011) was used to conduct a Wald chi-square test for the best fitting models.

Trials with response time less than 200 ms and above three standard deviations from a participant's mean response time were removed from the data (1.22% of responses). The factors used as fixed effects in the LME analysis included Response Time, Error Rate ("correct', "incorrect"), Instruction ("earlier", "later'), linear and quadratic component of Later-Probe Serial Position (serial position of the probe item that appeared later from the presented list), Distance (absolute value of the difference between two probe's serial positions), Intact/Reverse (whether probe order was consistent or inconsistent, with presentation order, respectively), Grouping ("grouped" vs. "ungrouped"), Within/Between group (whether the two probes were from the same group ("within group"), or from different groups ("between group"), Prediction Consistency (whether the prediction based on list-level position and group-level position is "consistent", "inconsistent", or "neutral"). The linear and quadratic component of the Later-Probe Serial Position were orthogonal to each other, generated with the "poly" function in R. We included the quadratic term to account for expected primacy and recency effect. Subject was included as a random effect on intercept. Error Rate, Instruction, Intact/Reverse, Grouping were treated as categorical factors. All other factors except Response Time were scaled and centered before being entered in the model. Response time was log-transformed to reduce skewness. The error rate data were fitted with logistic regression as it is a binary variable ("correct" vs. "incorrect"). LME estimated random effects first, followed by fixed effects. In the results tables, the "Estimate" column reported the corresponding regression coefficients, along with their standard errors. For the purposes of reporting the LME results, "incorrect" response, "intact" condition, "earlier" instruction,

"grouped", "within group" and "neutral" condition were set as the reference levels for the Error Rate, Intact/Reverse, Instruction, Grouping, Within/Between Group and Prediction Consistency factors, respectively. The best fits of LME models were obtained by conducting a series of iterative tests comparing progressively simpler models with more complex models using the Bayesian Information Criterion (BIC), as was done by Chapter 2 (Liu et al., 2014), using LMERConvenienceFunctions (Tremblay, 2013).

5.2.3 Results

We first analyzed the data involving both instructions ("earlier", "later") and both groups ("grouped", "ungrouped") to explore the effects of instructions and groups. Then we looked at "ungrouped" and "grouped" groups separately for the congruity effect. We then evaluated whether the data support the hierarchical positional coding model.

Effects of group and instruction

Figure 5.2 a,b,d,e and 5.3 a,b,d,e plot the error rate and response time measure as a function of both the earlier and later probe serial position for both instructions and both groups. A distance effect was evident on the bar plots across all conditions, with larger probe distance associated with lower error rates and faster response times. The "earlier"−"later" difference plots (Figure 5.2 c,f and Figure 5.3 c,f) show that instruction interacted with Probe serial positions, supported in the LME analysis by interaction between instruction and linear component of Later-Probe Serial Position (see Table 5.7, 5.9, 5.5, 5.10 for Figure 5.2 c,f and Figure 5.3 c,f respectively), and suggested this interaction was due to "earlier" instruction producing better performance at earlier serial positions, and "later" instruction producing better performance at later serial positions, in line with our predicted congruity effect. The "grouped"−"ungrouped" difference plots (Figure 5.2 g,h and Figure 5.3 g,h) did not show any clear pattern. Visual inspection also suggested the "grouped" group had less errors but slower response time than "ungrouped" (see Figure 5.1).

For both the error rate and response time measure, we included Instruction, Intact/Reverse, Distance, Grouping, linear and quadratic component of Later-Probe Serial Position as factors. The best error rate LME model with the lowest BIC value found a main effect of Instruction, Intact/Reverse, Later-Probe Serial Position, and a Instruction × Intact/Reverse interaction, but the main effect of Grouping was not significant. Because our fitting mecha-

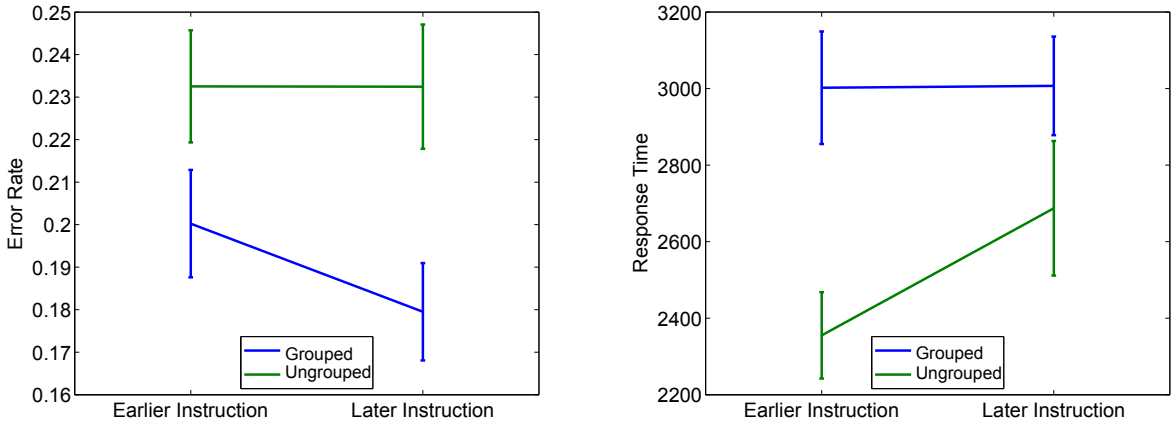


Figure 5.1: Main effect of error rate (left) and response time (right) for both groups and instructions.

	BIC	AIC	Log-likelihood	df
Best fitting LME model + Grouping	18656	18593	-9288.6	8
Best fitting LME model	18656	18602	-9293.7	7
Model difference ($\chi^2 = 10.381,\ p < 0.05$)	0	-9	5.1	1

Table 5.2: Model comparison of best fitting Error Rate model with the lowest BIC to the same model plus Grouping. Note that for BIC and AIC, lower numbers indicate better fit, but for log-likelihood, higher numbers indicate better fit. The log-likelihood ratio test using χ^2 test was significant.

nism returned a single best-fitting model with the lowest number of factors, we further added Grouping as a factor to the best-fitting LME model with lowest BIC value and compared it with the new model (summarized in Table 5.2). The new model with Grouping added fitted equally well as the original model using BIC, but with the AIC and log-likelihood measures of model fitness, the model that includes Grouping was prefered (Table 5.3). The best-fitting LME model with lowest BIC value showed "later" instruction, "ungrouped" group, and "reverse" presentation increased errors, whereas an increase in Distance and Later-Probe Serial Position reduced errors, confirming a distance effect, overall recency effect, and the main effect of Instruction and Grouping as shown in the left panel of Figure 5.1.

The best response time model fitted for BIC found a main effect of Distance, linear component of Later-Probe Serial Position, Grouping, quadratic component of Later-Probe Serial Position, and further found Instruction $\times$ linear component of Later-Probe Serial

		Estimate (SE)
Main effects		
	Intercept	-1.825(0.076)*
	Instruction	0.384(0.091)*
	Grouping	0.270(0.082)*
	Intact/Reverse	0.554(0.051)*
	Distance	-0.268(0.022)*
	Later serial position (linear)	-32.6(2.79)*
Interactions		
	Intact/Reverse × Instruction	-0.874(0.074)*

Table 5.3: The best-fitting LME model for error rate. The "Estimate" column reports the corresponding regression coefficient, along with its SE (standard error). Significant effects are denoted * - $p < 0.05$.

Position (Table 5.4). The LME model supported a distance effect, overall recency effect, a congruity effect and a U-shaped serial position effect.

Both the error rate and response time measure showed the benchmark effects. However, consistent with the prediction of positional coding models, grouping reduced error rate at a cost of increased response time. If one had only measured response time, one would have gotten only a partial picture of the effects of Grouping on behaviour due to an overall speed-accuracy tradeoff.

Congruity effect in "ungrouped" group

Analyzing both "grouped" and "ungrouped" groups together may not be the best approach to investigate the congruity effect, as we expected the "grouped" group may differ significantly from the "ungrouped" group in serial position effect and distance effect, and this could interfere with the congruity effect in unexpected ways. The congruity effect could be best analyzed by looking at both "ungrouped" and "grouped" group separately.

We expected to replicate the congruity effect reported in Chapter 2 (Liu et al., 2014) using a slightly longer list length. We divided the data by Grouping to further investigate the congruity effect (i.e., Instruction × linear component of Later-Probe Serial Position Interaction). The best-fitting model with the lowest BIC for error rates also showed a main effect of Distance, linear component of Later-Probe Serial Position. However, the congruity effect was not significant. To further examine if we could find a congruity effect,

		Estimate (SE)
Main effects		
	Intercept	7.788(0.047)*
	Instruction	-0.135(0.056)
	Intact/Reverse	0.066(0.010)
	Distance	-0.116(0.007)*
	Later(linear)	8.906(0.919)*
	Group	-0.199(0.055)*
	Error Rate	0.116(0.009)*
	Later(Quadratic)	-9.483(1.043)*
Interactions		
	Intact/Reverse × Instruction	-0.140(0.014)*
	Instruction × Distance	0.069(0.008)
	Instruction × Later(linear)	-22.30(1.134)*
	Distance × Later(linear)	2.994(0.847)*
	Error Rate × Distance	0.043(0.010)*
	Intact/Reverse ×Later(quadratic)	3.451(1.347)
	Instruction × Later(quadratic)	3.982(1.390)
	Intact/Reverse × Instruction × Later(quadratic)	-8.519(1.963)*

Table 5.4: The best-fitting LME model for response time. The congruity effect is in bold. The "Estimate" column reports the corresponding regression coefficient, along with its SE (standard error). Significant effects are denoted * - $p < 0.05$.

		Estimate (SE)
Main effects		
	Intercept	7.822(0.049)*
	Instruction	-0.095(0.071)
	Intact/Reverse	0.069(0.015)
	Distance	-0.127(0.010)*
	Later(linear)	11.87(1.449)*
	Later(Quadratic)	-6.328(0.755)*
Interactions		
	Intact/Reverse × Instruction	-0.166(0.021)*
	Intact/Reverse × Distance	-0.166(0.021)*
	Intact/Reverse × Later(linear)	-6.022(1.745)*
	Instruction × Distance	0.079(0.012)
	Instruction × Later(Linear)	-25.48(1.747)*

Table 5.5: The best-fitting LME model for "ungrouped" group's response time. The congruity effect is in bold. The "Estimate" column reports the corresponding regression coefficient, along with its SE (standard error). Significant effects are denoted * - $p < 0.05$.

we added Instruction × linear component of Later-Probe Serial Position interaction to the best-fitting model and found the congruity effect to be significant (Table 5.6). The model that included the congruity effect was reliably selected based on both AIC and log-likelihood (Table 5.8), thus we reported only the model that included the congruity effect. The best-fitting LME model for correct response time data showed a main effect of Distance, linear component of Later-Probe Serial Position and Instruction × linear component of Later-Probe Serial Position Interaction, supported a significant distance effect, serial position effect and congruity effect respectively (Table 5.5). In sum, the "ungrouped" group replicated results from Chapter 2 (Liu et al., 2014), supporting the robustness of distance effect, recency effect and congruity effect.

Congruity effect in "grouped" group

We first analyzed the "grouped" group using the same factors as in the "ungrouped" group. The best-fitting LME model with lowest BIC did not show a significant congruity effect on error rate. We refitted the model using AIC and found a significant congruity effect (Table 5.9); however, the best-fitting model with the lowest AIC had a higher BIC than the best-fitting LME model, thus the results should be interpreted with caution. A congruity

		Estimate (SE)
Main effects		
	Intercept	-1.523(0.088)*
	Instruction	0.430(0.127)*
	Intact/Reverse	0.437(0.069)*
	Distance	-0.299(0.031)*
	Later serial position (linear)	-0.224(0.028)*
Interactions		
	Intact/Reverse × Instruction	-0.857(0.102)*

Table 5.6: The best-fitting LME model for "ungrouped" group's error rate. The "Estimate" column reports the corresponding regression coefficient, along with its SE (standard error). Significant effects are denoted * - $p < 0.05$.

		Estimate (SE)
Main effects		
	Intercept	-1.512(0.088)*
	Instruction	0.409(0.127)*
	Intact/Reverse	0.434(0.069)*
	Distance	-0.300(0.031)*
	Later serial position (linear)	-24.07(4.951)*
Interactions		
	Intact/Reverse × Instruction	-0.857(0.102)*
	Instruction × Later serial position (Linear)	-15.02(6.797)*

Table 5.7: The best-fitting LME model for "ungrouped" group error rate with addition of the congruity effect . The congruity effect is in bold. The "Estimate" column reports the corresponding regression coefficient, along with its SE (standard error). Significant effects are denoted * - $p < 0.05$.

	BIC	AIC	Log-likelihood	df
Best fitting LME model + Congruity effect	9718.1	9660.9	-4822.4	8
Best fitting LME model	9713.8	9663.7	-4824.9	7
Model difference ($\chi^2 = 4.864,\ p < 0.05$)	4.3	-2.8	0.5	1

Table 5.8: Model comparison of best fitting LME model with the lowest BIC to the same model plus Instruction × linear component of Later-Probe Serial Position. Note that for BIC and AIC, lower numbers indicate better fit, but for log-likelihood, higher numbers indicate better fit. The log-likelihood ratio test using χ^2 test was significant.

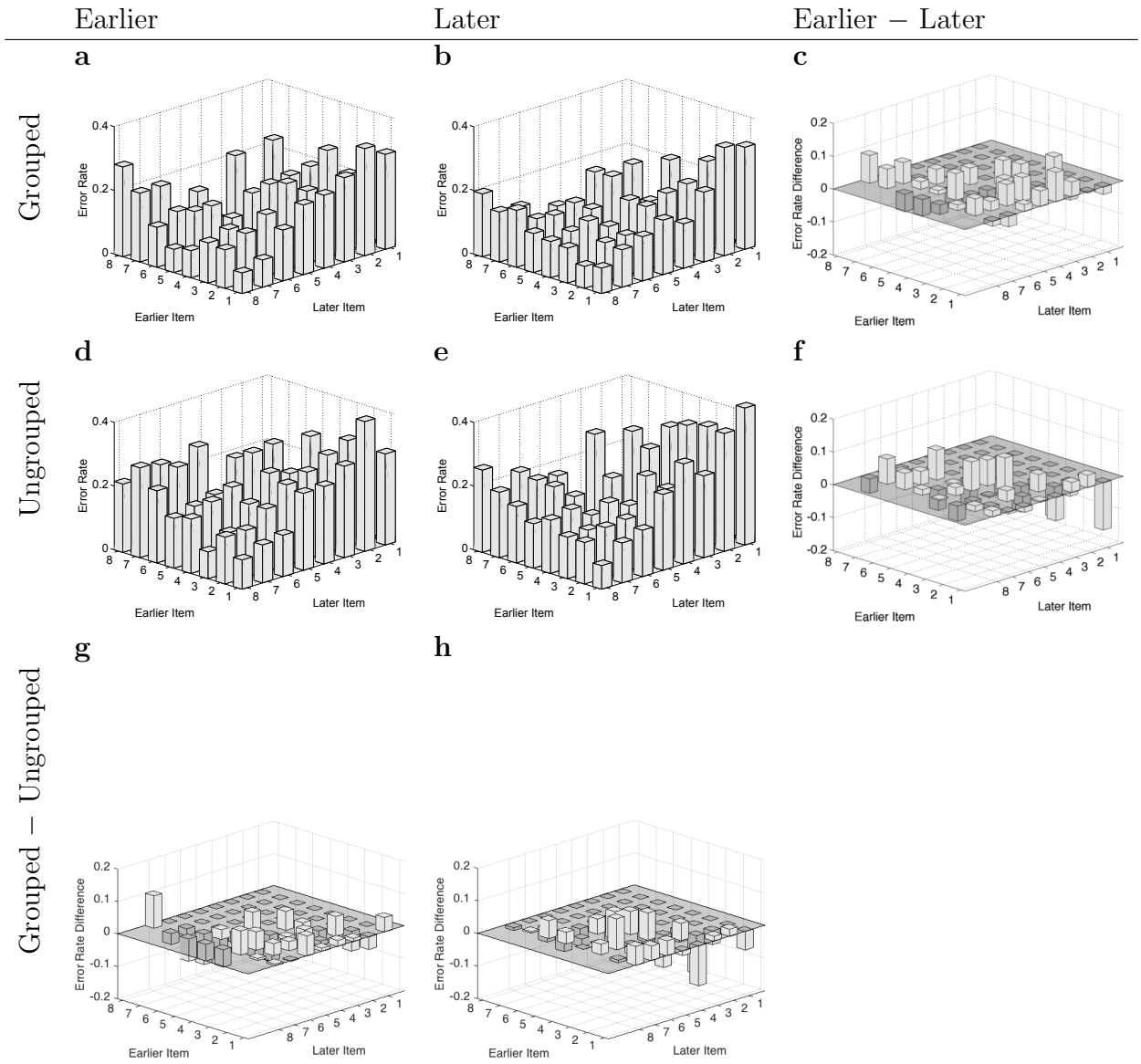


Figure 5.2: Error rate as a function of both probe items' serial position (earlier item and later item, respectively) broken down by groups ("grouped", "ungrouped", and the difference "grouped" – "ungrouped"), and instruction ("earlier", "later" and the difference, "earlier"−"later") in columns. The differences plots are corrected for mean error rates.

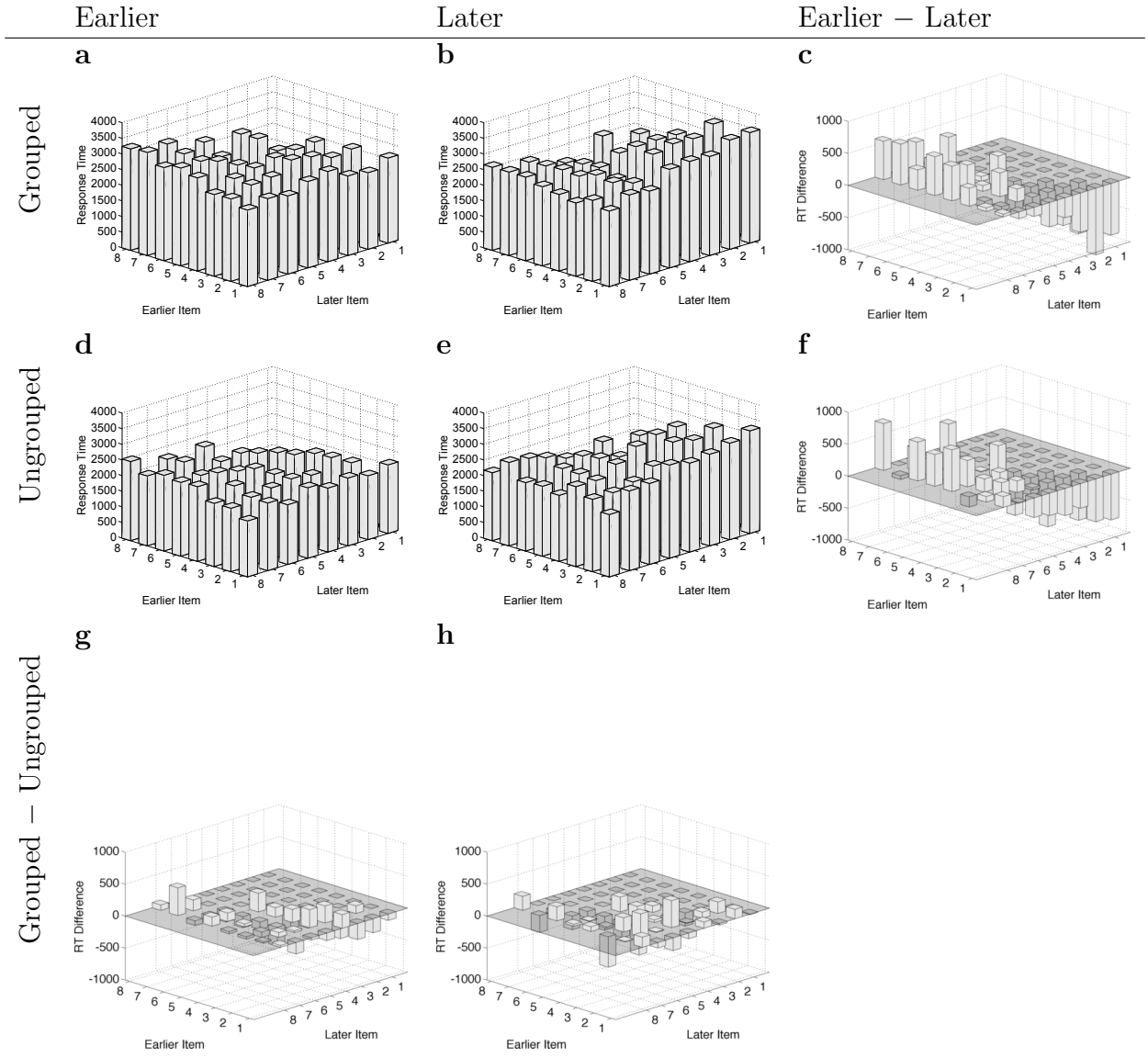


Figure 5.3: Response time as a function of both probe items' serial position (earlier item and later item, respectively) broken down by groups ("grouped", "ungrouped", and the difference "grouped" – "ungrouped"), and instruction ("earlier", "later" and the difference, "earlier"−"later") in columns. The differences plots are corrected for mean response times.

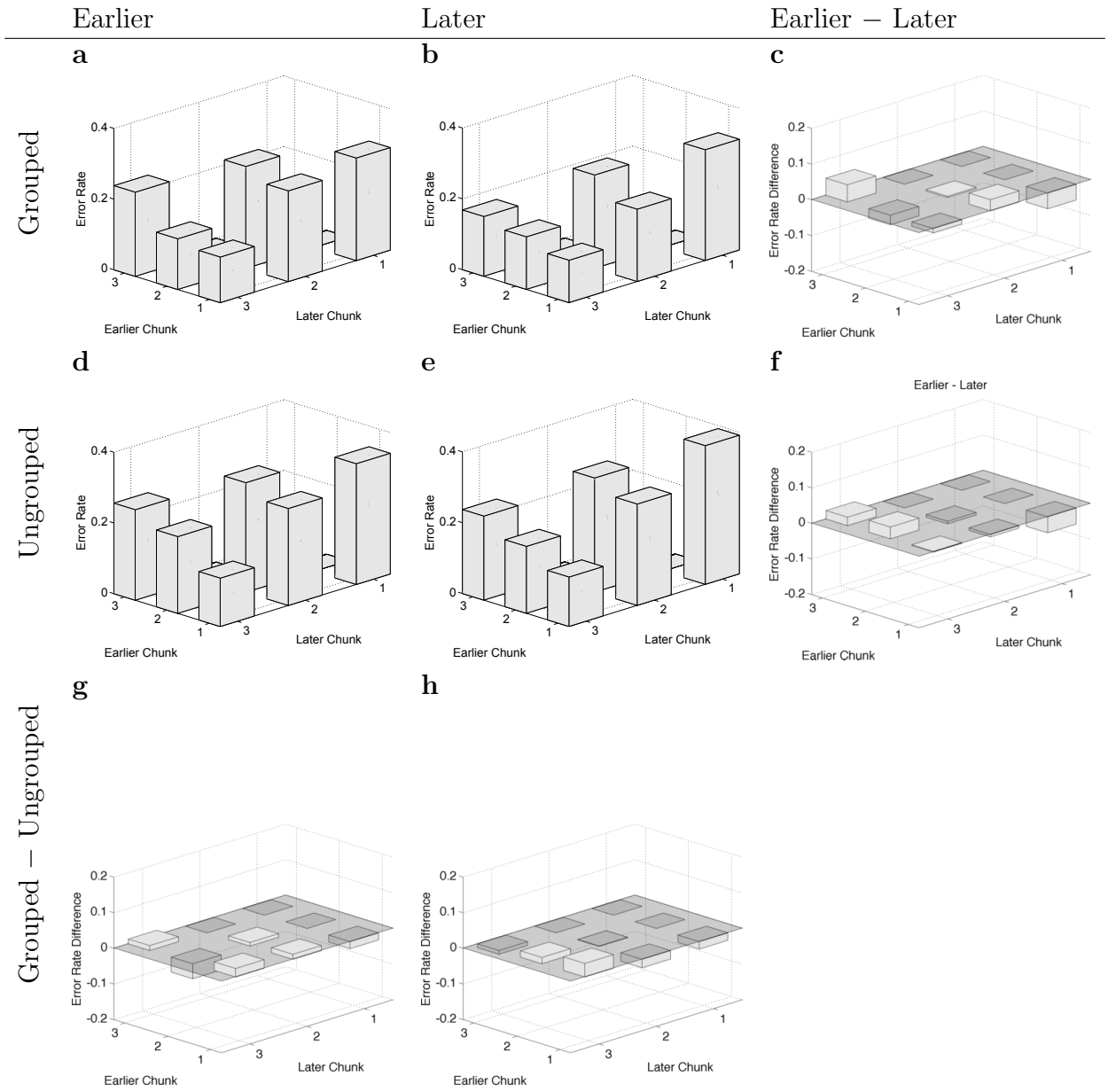


Figure 5.4: Error rate as a function of both probe items' group position (earlier group and later group, respectively) broken down by groups ("grouped", "ungrouped", and the difference "grouped" – "ungrouped"), and instruction ("earlier", "later" and the difference, "earlier"−"later") in columns. The differences plots are corrected for mean error rates.

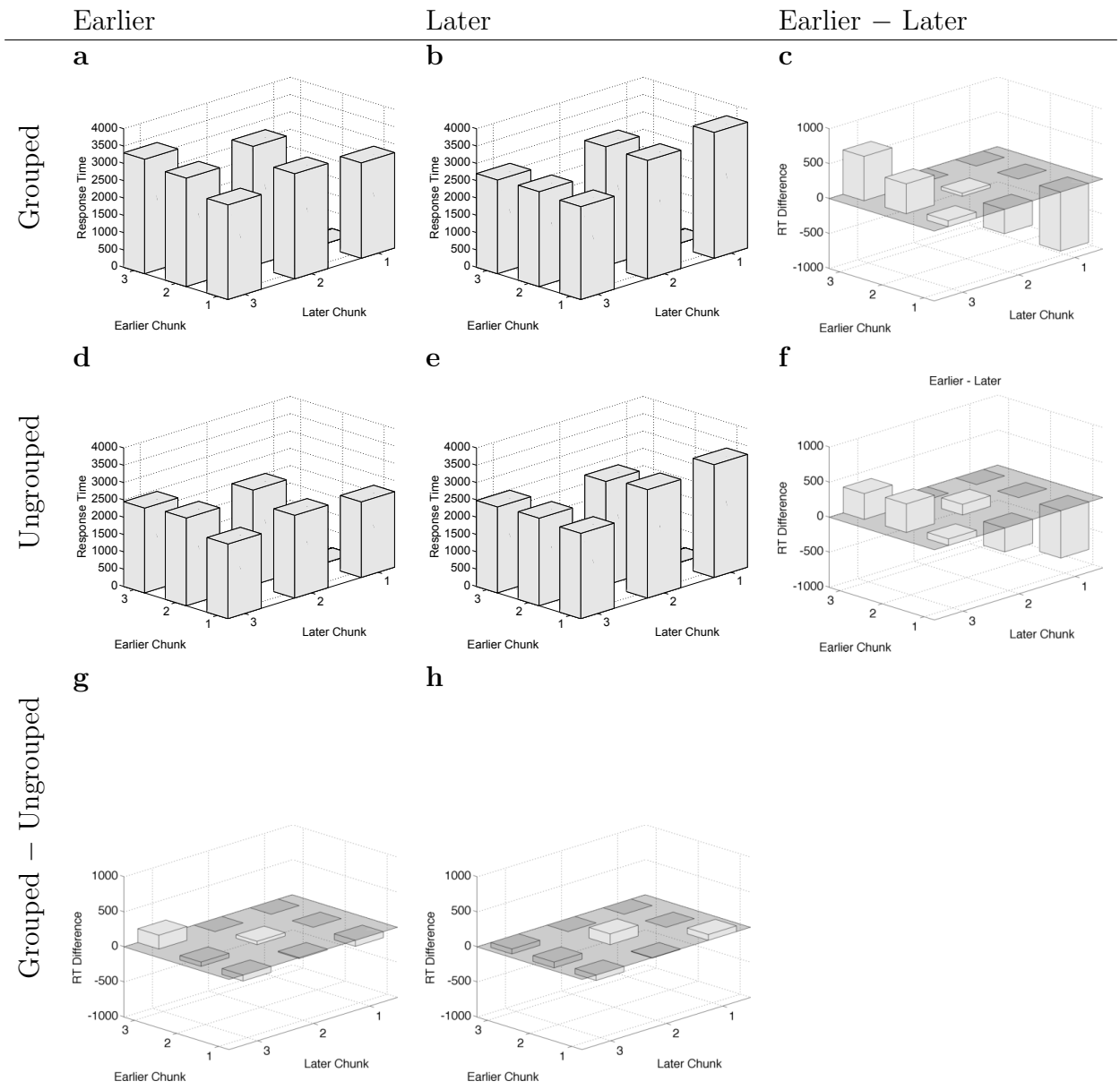


Figure 5.5: Response time as a function of both probe items' group position (earlier group and later group, respectively) broken down by groups ("grouped", "ungrouped", and the difference "grouped" – "ungrouped"), and instruction ("earlier", "later" and the difference, "earlier"−"later") in columns. The differences plots are corrected for main response times.

		Estimate (SE)
Main effects		
	Intercept	-1.867(0.092)*
	Instruction	0.345(0.131)*
	Intact/Reverse	0.687(0.074)*
	Distance	-0.233(0.033)*
	Later serial position (linear)	-34.76(5.271)*
Interactions		
	Intact/Reverse × Instruction	-0.896(0.108)*
	Instruction × Later serial position (Linear)	0.754(7.140)*

Table 5.9: The best-fitting LME model for "grouped" group error rate plus the congruity effect. The congruity effect is in bold. The "Estimate" column reports the corresponding regression coefficient, along with its SE (standard error). Significant effects are denoted * - $p < 0.05$.

effect was found for "grouped" group response time, supported by a significant Instruction × linear component of Later probe serial position interaction in the best-fitting correct response time model (Table 5.10). The congruity effect was characterized by a positive Later probe serial position slope for "earlier" instruction and negative slope for "later" instruction. The "ungrouped" groups showed a significant main effect of Instruction, with the "earlier" instruction participants responding faster than the "later" instruction, but the "grouped" group did not show a main effect of Instruction. Note that the error-rate congruity effect was small compared with the response time congruity effect, and we may not have a big enough sample size in both groups to test its interaction with grouping. The results suggested grouping did influence order judgement speed, but we found no evidence it interacted with the congruity effect.

Comparing hypotheses derived from comparative judgement and positional coding theory

According to the comparative judgement literature (Holyoak & Mah, 1981), a distance effect is indicative of magnitude comparison. A between-group judgement was thought to not require a magnitude comparison of the individual items, as the use of group labels were sufficient for making the judgement. Therefore, between-group judgements should show no distance effect or reduced distance effect than within-group judgements. A null finding of

		Estimate (SE)
Main effects		
	Intercept	7.823(0.049)*
	Instruction	0.095(0.071)
	Intact/Reverse	0.069(0.014)
	Distance	-0.127(0.010)*
	Later(linear)	11.88(1.450)*
	Later(Quadratic)	-6.328(0.755)*
Interactions		
	Intact/Reverse × Instruction	-0.166(0.021)*
	Intact/Reverse × Distance	0.044(0.012)*
	Intact/Reverse × Later(linear)	-0.602(1.745)*
	Instruction × Distance	0.079(0.012)
	Instruction × Later(linear)	-25.48(1.747)*

Table 5.10: The best-fitting LME model for "grouped" group response time. The congruity effect is in bold. The "Estimate" column reports the corresponding regression coefficient, along with its SE (standard error). Significant effects are denoted * - $p < 0.05$.

between-group distance effect would suggest the group labels were nominal, whereas a finding of reduced distance effect would be inconsistent with the nominal representation of group labels. As a consequence of reduced distance effect, the response for "grouped" lists should be less accurate and slower than "ungrouped" lists. This was different from the prediction from two-level positional coding model, where it predicted the "grouped" list would had higher accuracy at the expense of slower response time.

Our JOR data are inconsistent with predictions from some theories of comparative judgement; however, they were compatible with the two-level positional coding model. Following this line of thought, we further plotted cumulative density functions of the response time distributions, collapsing all participants together, and found the "grouped" list performed slower than the "ungrouped" list across a very large percentile spectrum of the cumulative density function, a relationship of approximate stochastic dominance (Figure 5.6). This result further supported the idea that there were two levels of position codes and that additional processing was required for the use of group-level codes. Note that in serial recall, grouping produced faster response times rather than slower response times found in JOR. However, making responses in serial recall may not require the additional task of making a comparison between two items. The differential response times pattern between JOR and

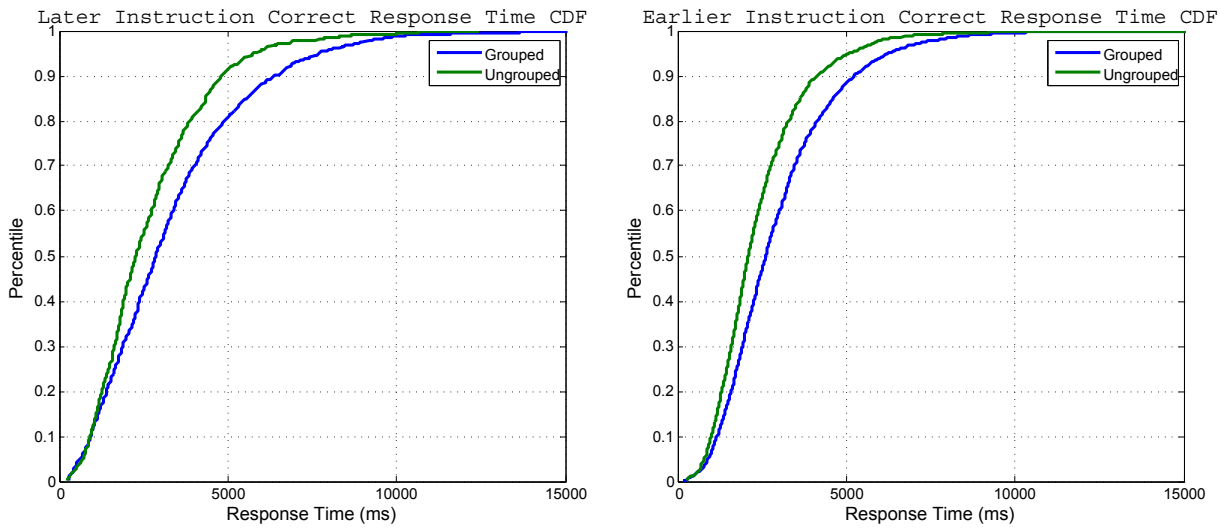


Figure 5.6: Cumulative density functions of "grouped" group response time by the Earlier (right) and Later instruction (left).

serial recall could be due to the underlying comparison mechanism specific to JOR and comparative judgements.

To directly test the Group Distance × Grouping and Within/Between Group × Grouping interactions, we constructed a LME model with seven fixed factors (Instruction, Grouping, Intact/Reverse, Group Position, Group Distance, Response Time, Within/Between Group) and with two-way interactions of Group Distance × Grouping, Within/Between Group × Grouping. The group position and group distance were analogous to the list-level serial position effect and distance effect, but used group-level position codes instead of list-level position codes. The Group Distance was calculated by the absolute value of the Group Code differences. The Within/Between Group factor was coded for whether the two probes were from the same group ("within group"), or from different groups ("between group"). The group position and group distance were continuous factors, while the Within/Between Group was a categorical factor with "between group" as the default level. The LME analysis summarized in Table 5.11 found the "reverse" presentation, lower Later Group Position, smaller Group Distance, the "ungrouped" group, and faster response time predicted higher error rates than the "intact" presentation, larger Later Group Position, larger Group Distance, the "grouped"

group, slower responses, respectively. The main effect of Within/Between Group was significant, showing within-group judgements had lower errors; however, this effect was one order of magnitude smaller than other main effects, and this factor needed to be understood through its interaction with Grouping. Turning to the interactions, the Grouping × Group Distance interaction was significant, supporting the idea that the "ungrouped" group had a larger distance effect than the "grouped" group. The Grouping × Within/Between group interaction was marginal, not significant ($p<0.1$). We refitted the model by using the "within" level as reference, and further confirmed between-group comparison was more accurate in the "grouped" group than in the "ungrouped" group. Our results were consistent with the finding that grouping reduced between-group distance effect in comparative judgements (Howard, 1980; Kosslyn et al., 1977; Maki, 1982; Woocher et al., 1978). A significant group label distance effect may suggest group labels were not nominal, with the caveat that for a list with three groups, the biggest distance also happened to include the first and last group, which should be better than the middle group due to the bow-shaped serial position effect. We also replicated previous findings that between-group comparisons were easier than within-group comparisons for the "grouped" group, and further showed this effect reverses in the "ungrouped" group.

So far we have shown that the JOR grouping data were consistent with the two-level hierarchical positional coding model in predicting overall error rate and response time patterns, the distance effect patterns and within/between-group error rate patterns. We further test the prediction-consistency predictions proposed in the introduction section by adding Prediction Consistency (whether the prediction based on list-level position and group-position was consistent, inconsistent, or neutral) as a fixed effect to the model in Table 5.11, with the reference level set to "neutral". The new model, presented in Table 5.12, further showed the "consistent" condition was more accurate than the "neutral" condition, and the "neutral" condition was more accurate than the "inconsistent" condition. In the LME model (Table 5.12) the prediction consistency effect slopes were larger than the main effect of grouping, and the main effect of grouping became non-significant, further suggesting positional coding could be a main mechanism underlying the grouping effect.

		Estimate (SE)
Main effects		
	Intercept	-5.095(0.296)*
	Instruction	-0.104(0.091)
	Intact/Reverse	0.149(0.037)*
	Later Group Position	-0.224(0.021)*
	Grouping	0.417(0.098)*
	Group Distance	-0.190(0.053)*
	Within/Between Group	-0.037(0.109)*
	Response Time	0.446(0.036)*
Interactions		
	Grouping × Group Distance	-0.145(0.071)*
	Grouping × Within/Between Group	-0.279(0.150)·

Table 5.11: LME model testing grouping effects. The "Estimate" column reports the corresponding regression coefficient, along with its SE (standard error). Significant effects are denoted * - $p < 0.05$ and · - $p < 0.1$

		Estimate (SE)
Main effects		
	Intercept	-4.852(0.310)*
	Instruction	-0.103(0.091)
	Intact/Reverse	0.149(0.037)*
	Later Group Position	-0.224(0.021)*
	Grouping	0.142(0.147)
	Group Distance	-0.194(0.053)*
	Within/Between Group	-0.136(0.114)
	Response Time	0.430(0.036)*
	Prediction Consistency (Consistent)	-0.161(0.056) *
	Prediction Consistency (Inconsistent)	0.196(0.053) *
Interactions		
	Grouping × Group Distance	-0.145(0.071)*
	Grouping × Within/Between Group	-0.271(0.150)·

Table 5.12: LME model testing grouping effects. The "Estimate" column reports the corresponding regression coefficient, along with its SE (standard error). Significant effects are denoted * - $p < 0.05$ and · - $p < 0.1$

5.3 Discussion

In this experiment, we found the benchmark effects of comparative judgements, the serial position effect, distance effect and the congruity effect all generalized to the grouped lists, consistent with previous findings on the comparative judgement paradigm (e.g., Jou, 2010). We further showed the benchmark effects could be found on the error rate measure. We found no clear evidence of grouping interacting with serial position effects. Taken together, we supported the generality of the congruity effect. Although grouped sets have been a major topic of interest in comparative-judgement research, those studies nearly always use only two groups and the groups are usually category labels which are considered to be nominal (e.g., buildings versus animals). The grouping studies in serial recall typically use more than two groups, and the group labels are considered to be ordinal (e.g., group 1, group 2 and group 3). Our use of three groups allowed us to further ask whether group labels can carry order information and use a grouping method consistent with serial recall grouping studies. If the group labels were nominal, we would expect between-group judgement speed to be the same. We found between-group judgements were faster and more accurate as group distance increases. This result suggests group labels could be used for JORs and the labels are ordinal rather than nominal.

Response-time data from the comparative judgement paradigm generally shows the serial position curve is a smooth single bow shape, instead of a "scalloped" pattern as found in serial-recall serial-position curves, and this has been taken as support that comparative judgement tasks act on a one-dimensional memory representation (Jou, 2011). Jou (2011) further suggests the lack of grouping effects could be caused by ambiguous effect of using group boundary items as references. We suggest the lack of "scalloped" serial position pattern may not be the best indicator of a lack of grouping strategy. In comparative judgements, both probes need to be considered for serial position, and the distance effect and congruity effect may have unknown effect on the serial position curves. Further, positional coding models could assume the comparative judgement can compare the group-level code directly, as an alternative to use of boundary item as references. In this study, we do find converging evidence to support hierarchical memory representations. First, Jou (2011) suggests the task difference between comparative judgement and serial recall is responsible for how lists are represented and processed, where the comparative judgement tasks show no evidence of hier-

archical representation in contrast to what is found for serial recall. We suggest caution when comparing response time results from comparative judgements to the accuracy results from serial recall. A fast response time could reflect either an easier judgement, or a more difficult judgement but the participant made a quick guess (Petrusic, 1992). This is supported by our finding that grouping enhances overall judgement accuracy but increases response time. The accuracy finding could be accounted for by two-level hierarchical positional coding models, where the two-level position code has redundancies from both the group-level and list-level position codes that enhance information quality. The two-level hierarchical positional coding models could also explain the increased response times if we assume processing of the added dimension of position codes takes additional time. See below where we elaborate more on the evidence for hierarchical positional coding and evaluate other alternative models.

In sum, there is evidence to support the idea that group labels may carry order information, and controlling speed-accuracy tradeoff may allow direct comparison between the serial recall and comparative judgements results. We further show the JOR error rate results are compatible with two-level positional coding models.

5.3.1 Relating to the serial recall paradigm

Because the comparative judgement task and the serial recall task both test order memory, one would expect consistent evidence from both tasks. However, comparative judgement studies and serial recall have been studied and theories have been developed independently. Not until recently have the two paradigms been evaluated together (e.g., Jou, 2011). Results from comparative judgement tasks favour an one-dimension position code (Banks, White, & Mermelstein, 1980; Bower, 1971; Jou, 2005, 2011; Woocher et al., 1978), whereas the serial recall results favour a hierarchical position code (Henson, 1998; Lee & Estes, 1981; Maybery et al., 2002; Ng & Maybery, 2002, 2005; Wickelgren, 1967; Ryan, 1969a). The differences between the two tasks have been attributed to differential demands of the tasks, where comparative judgements test for relative order information and serial recall test for absolute order information (Jou, 2011). In other words, comparative judgements do not require retrieval of the intervening item from the list, whereas serial recall requires retrieval of all items from the list. There is evidence that this might be true, as researchers found JOR performance could not be predicted from absolute position information of both probes (Hacker, 1980; but see Lockhart, 1969), and the serial position curves for comparative judgements were

different from serial recall (Jou, 2011). Here we suggest that despite the differences between the tasks, it may not be straightforward to compare response time results from comparative judgements to error rate results from serial recall, as speed-accuracy tradeoffs may differ between the two tasks. We found grouping reduces errors for both the JOR and serial recall, yet grouping increases response time for the JOR task and reduces response time for the serial recall task. Therefore the use of error rate data in comparative judgements is essential to evaluate whether this relative-order task challenges the use of hierarchical position codes. In this experiment, we found the JOR accuracy is enhanced by grouping. This finding is consistent with the prediction of two-level hierarchical position codes, where redundancy from position codes enhances order memory. We further show that when within-group position codes predict different results from the list-level position codes, participants made more errors, and when within-group position codes predict the same results from the list-level position codes, participants made fewer errors. In sum, the error rate patterns suggest despite the task differences, error rate results from comparative judgements might be compatible with a hierarchical position codes. In addition, when group labels are conceptualized as superordinate position code, the hierarchical positional coding models could be extended to explain results from comparative judgements.

Although positional coding models could explain the general pattern of comparative judgements, positional coding models need additional assumptions to explain the response time pattern. First, we hypothesize processing two-dimensional position codes takes longer time than processing one-dimensional position codes, and cannot be faster than processing one-dimensional position codes, as the judgement could use extra sources of information from the second dimension. The stochastic dominance of the grouped lists response times supported this hypothesis. Alternatively, theories from the comparative judgement literature could also explain speed-accuracy tradeoffs by evidence accrual models, where information accumulates towards reaching a decision (Petrusic, 1992).

5.3.2 Direct access models and chaining models

Direct access models assume that item positions could be accessed directly from the list. A direct access model typically uses position codes, or order codes, where the position code could be used to access the associated item directly. For example, SIMPLE (Brown et al., 2007) and OSCAR (Brown et al., 2000) uses a temporal code, where memory are represented

along the temporal dimension, whereas Farrell (2012b) proposed a model uses an absolute position code. As presented in Chapter 2.7, the SIMPLE model could account for the congruity effect on supraspan lists by adding a bias gradient to the temporal dimension; However, it could not account for subspan JOR results. Because we induced grouping by manipulating the temporal dimension, the SIMPLE model might be able to fit our data by adding temporal gaps to the temporal dimension, and no additional mechanism may be required to fit the grouping effect. Alternatively, it is also possible that SIMPLE could fit the grouping results by adding a group-level position code using absolute position codes. However, we cannot confirm either account without further analysis. It is worth noting that SIMPLE cannot explain serial recall grouping results, as output direction dominates the recall serial position curve (see Chapter 4). Data from the current experiment also suggests absolute position coding could account for the grouping effects, where the Prediction Consistency factor shows the consistent condition is more accurate than the neutral condition and the neutral condition is more accurate than the inconsistent condition. However, to account for the JOR benchmark effects, more assumptions need to be made. For example, the congruity effect could possibly originate from serial recall strategies, where the "earlier" instruction tapping forward serial recall mechanisms and the "later" instruction tapping backward serial recall mechanisms. This is further supported in Chapter 4 by mirror images of forward and backward serial position curves. It is possible a combination of temporal and absolute position coding mechanisms could account for the full spectrum of the JOR effects.

A chaining model assumes order information is stored by forming associations between successive items of a sequence, where retrieval of order information requires probing with the first item of the chain to retrieve the full list through item-to-item association (Caplan, 2015). To account for missing items, models are developed to use remote associations as a cue, in addition to direct item pairwise association (e.g., TODAM; Lewandowsky & Murdock, 1989). The chaining models could explain subspan JOR results by start cueing from the beginning of the list for the "earlier" instruction, versus start cueing from the end of the list for the "later" instruction. However, it remains a challenge for the chaining models to account for distance effects found in supraspan JORs, as well as to account for the congruity effect and grouping effects. A more elaborated chaining model could potentially account for the JOR grouping results, which need further future modelling work.

In summary, our results suggest the congruity effect can be found regardless of grouping

of the list, and when the error rate measure is used, comparative judgement and serial recall both support hierarchical positional coding .

Chapter 6

Discussion

The work presented in the earlier chapters collectively helps us to better understand a specific question that has broader implications for both memory and comparative-judgement research: Is the congruity effect is a general characteristic of order memory? The answer to this question helps us to advance the fields of order memory and comparative judgements, and to provide constraints for existing models. In this final chapter, we discuss the theoretical implications of the findings presented in the earlier chapters, and discuss what new knowledge have been gained from the sets of results. The discussion is divided in three sections. Each section addresses important concepts from this dissertation. First we will discuses the benchmark effects of JOR, including the serial position effect, distance effect and congruity effect. In the next section we discuss how we establish the theoretical link among the JOR paradigm, comparative judgements and serial recall. The next section is dedicated to discuss how the findings constrain memory models and how to further develop existing memory models.Then we discuss the limitations of the current work, and finally we discuss future directions following this work.

6.1 Benchmark effects of JOR's

In Chapter 1, we introduced the three JOR benchmark effects: a) a serial position effect or edge effect, characterized by faster and more accurate responses for items at the beginning or end of the list compared with items in the middle of the list; b) a distance effect, characterized by faster and more accurate responses for probes when the difference between probe positions increase; and c) a congruity effect, characterized by facilitation of response times and accuracy when the instruction is congruent with the probe item positions, and

this effect reverses when the instruction is not congruent with the probe item positions. The congruity effect of the JOR task was first reported by Chan et al. (2009), testing subspan lists of consonants; thus, it is fairly recent and had not been systematically investigated; this motivated most of the research contained within this dissertation. We now consider each of these effects in turn.

6.1.1 Serial position effects

The enhancement of response time at the edges of the list has been found in our studies, and we show this serial position effect generalizes across supraspan lists of words and consonants (Chapter 2, Liu et al., 2014), the 26-item English alphabet and to lists that are grouped into sub-lists (Chapter 4). Our work replicated previous studies using the recency judgement paradigm (Hacker, 1980; Muter, 1979; Yntema & Trask, 1963) by asking "which item came later?". We have assumed the "later" and "recent" instructions are asking the same question; however, asking "which item is more recent?" implies a temporal relationship of probes, where the reference point is the current time. Asking "which item came later" does not necessarily imply a temporal relationship and nor a specific reference point. For the "later" instruction, the relationship between probe items could be rank order instead of temporal order, and the reference point could be anywhere from the list. For example, for judging relative order of letters from the English alphabet, it does not make sense to ask which item is more recent. It is possible that this minor difference between the "later" and "recent" instruction could lead to enhanced recency effects from the "recency" instruction than "later" instruction, because the "recent" instruction may induce the current time as a reference, which is closer to more recent list items. We need future studies to test for this hypothesis directly.

The generality of the response time serial position effects on supraspan lists suggests this benchmark effect could be used to constrain memory theories. The results from response time serial position effects and error rate serial position effects should be interpreted with caution, as different mechanisms may underly the two measures, we will further address this topic in the second section of the chapter.

6.1.2 Distance effect

The distance effect is another robust effect found in comparative judgement studies (e.g., Banks, 1977; Moyer & Bayer, 1976; Jou, 2011), but not widely reported in JOR studies using the "recency" instruction prior to the studies in this dissertation (e.g., Hacker, 1980; Muter, 1979). The lack of report of distance effect maybe due to the lack of connection between the two fields of research. We replicated the distance effect using both the response time and error rate measure using different procedures, and showed that the distance effect is found regardless of the list materials (Chapter 2, Liu et al., 2014), episodic/semantic memory (Chapter 3) and grouping (Chapter 5). However, our results suggest the distance effect is modulated by the grouping structure of the list. The distance effect is larger for between-group probes than within-group probes for the grouped lists.

The comparative judgement paradigm has yielded mixed results about the effects of grouping on the distance effect. The majority of the grouping studies have used only two groups, or categories. The distance effects are reduced (Maki, 1981) or disappear (Kosslyn et al., 1977; Pohl, 1990; Pliske & Smith, 1979) for judgements between groups. The consensus is that the distance effect reflects a magnitude comparison process and the disappearance of distance effect is because when group labels could be used for judgements, no magnitude comparison is required (Schweickart & Brown, 2013). For the positional coding models, the group labels still carry magnitude information, predicting a distance effect at the group level. In this regard, our findings are consistent with a hierarchal positional coding explanation, where group labels as well as within-group position codes could be used for relative order judgements. We further found a distance effect; at the group level, as group distance increases, the error rates decreases.

6.1.3 Congruity effect

Although we replicated the serial position effect and distance effect, the main focus of this dissertation is on the congruity effect, because it has been overlooked by episodic memory research. The serial position effects are already well established in the JOR paradigm (Chan et al., 2009; Hacker, 1980), and the distance effect is well established in comparative judgements (Banks, 1977; Moyer & Bayer, 1976; Jou, 2011); however, the congruity effect in supraspan lists was not reported before the work of Chapter 2 (Liu et al., 2014), Chap-

ter 3 and Chapter 5. Our results demonstrate the congruity effect could be found using both response time and error rate as measures, coexisting with the serial position effect and distance effect, and this effect could be found to generalize across grouping structure, semantic/episodic memory. The generality of the congruity effect suggests the current versions of order memory models need further assumptions to account for this effect. In Chapter 2 (Liu et al., 2014), we fitted Hacker's self-terminating search model to data and found that assuming forward search direction for the "earlier" instruction and backward search direction for "later" instruction could only account for subspan JOR results, where the "earlier" instruction resembles forward self-terminating search and "later" instruction resembles backward self-terminating search behavioural patterns. For the supraspan JOR, a congruity effect is found independent from an overall recency effect, and the behavioural pattern can no longer be explained by scan direction reversal. As a follow-up analysis, we fitted SIMPLE to the data from Chapter 2, assuming a correct response is made when the retrieved item positions are in correct relative order (see Chapter 2.7). We found that the unmodified SIMPLE could not account for the congruity effect. However, we found that adding a bias gradient to the temporal code that either selectively increases temporal discriminability toward the beginning of the list or toward the end of the list could capture the congruity effect of supraspan lists qualitatively. We will further discuss this in the modelling section.

6.1.4 Subspan versus supraspan lists

Recall that in the introduction we explained that we do not use the terms "subspan" and "supraspan" to refer to the construct of memory span itself. Rather, we use these terms to distinguish short lists where response accuracy is at ceiling, from long lists where response accuracy is below ceiling, respectively. In Chapter 2 (Liu et al., 2014), we tested both subspan (list length 4, consonants) and supraspan (list length 8 consonants, list length 4, 6, 8, 10 nouns). Subspan JORs do not show both primacy and recency effect, characterized by a significant quadratic component on the serial position curve. Instead, the "earlier" instruction produces a primacy effect without a recency effect and the "later" instruction produces a recency effect without a primacy effect. This serial position pattern is consistent with a forward self-terminating search and backward self-terminating search behavioural pattern, for "earlier" and "later" instruction, respectively. Chan et al. (2009) did not look for a distance effect. In Chapter 2 (Liu et al., 2014), with a greater sample size than Chan

et al.'s (2009) study, we show a distance effect can be found for subspan lists. The finding of a distance effect in subspan lists suggests although the distance effect in subspan list might be smaller in magnitude and explains less variance than in supraspan lists, JORs may share some underlying mechanism. We also found the congruity effect to be significant for both subspan and supraspan lists for response times, but for the error rate, the congruity effect was not significant in subspan lists. The differential findings between the response time and error rate measures could be partially attributed to the accuracy being at ceiling; thus, the error rate measure may not be a sensitive measure. To summarize, the results suggest a common mechanism may underly both the distance effect and the congruity effect for both subspan and supraspan lists. However, successful fit of Hacker's model to the congruity effect of subspan lists but not congruity effect of supraspan lists, and successful fit of a slightly modified SIMPLE model to the error rate congruity effect of supraspan list but not the congruity effect of subspan lists, suggested there might be differences in the underlying mechanisms. We suggest this underlying difference between subspan and supraspan lists is independent from the distance effect and the congruity effect. Our data cannot rule out models that could predict the distance effect and the congruity effect regardless of list length, and could also account for the subspan serial position effects. We will revisit this topic in the model discussion section.

6.2 Comparing order memory tests

Order memory is broadly defined as the capacity to remember order of items from a specified sequence. By this definition, the JOR task, serial recall task, and comparative judgement task are all order memory tests. In this section we further discuss how those tasks are related.

6.2.1 JOR as comparative judgement

The comparative judgement task asks participants to make relative magnitude judgement of two items along a continuous dimension (Moyer & Bayer, 1976; Banks, 1977; Petrusic, 1992; Jou, 2011). Researchers used the comparative judgement paradigm to study how people make order judgements (Jou & Aldridge, 1999; Jou, 2003, 2011), with a focus on well learned lists and the response time measure. According to this definition, the JOR task can be viewed as a subtype of comparative judgement where probe order is judged along the

dimension of time. Our results supported this position, in that the JOR paradigm shares the same benchmark effects found with the comparative judgement paradigm, and this has been repeatedly demonstrated across list lengths, testing materials, and grouping organization of the lists. However, the JOR congruity effect is not identical to congruity effects found in the comparative judgement paradigm. In the comparative judgement paradigm, the congruity effect is the interaction between semantic properties of items and the instruction. In the JOR paradigm, congruity effect is the interaction between item serial position and instruction. The semantic properties of items are most of time ordinal and the interval is not defined precisely (e.g., it is quantifiable how "large" is different from "small"), unlike the temporal dimension. Our findings from the JOR studies suggest order memory along the temporal dimension may not be different from order memory along any continuum, and a unitary theoretical account may underly all order memory phenomena.

6.2.2 Insights from the grouping results

We focused on JORs as a means of testing memory for order. However, order memory has been extensively studied using the serial recall paradigm. In serial recall, participants need to remember both the items and their order relative to the list; however, in JORs, strong item memory may not be essential, as the items are given to participants, and participants only need to remember the order of one item relative to another item. The difference between JORs and serial recall parallels the argument made for how comparative judgements may differ from serial recall based on results of grouping studies, as the JOR task could be considered as a special case of the comparative judgement task.

The results from the serial recall paradigm found clear evidence of grouping on the serial position effects (see Chapter 4). However, studies using the comparative judgement paradigm typical found grouping does not affect the serial position function, unless the group structure is very easy to use (Jou, 2011). For instance, when the list structure is trained to a very high criterion (Kosslyn et al., 1977), when the list structure is already from pre-existing groups or semantic categories (e.g., buildings versus animals; Maki, 1981; Pliske & Smith, 1979) and when the group relation is also true in the real world (e.g., buildings are usually larger than animals in the real world) participants could use grouping to facilitate judgements. Jou (2011) suggests the different findings of grouping effects on the serial recall and comparative judgement paradigm are because of the different use of local reference points. Jou (2011)

argues that local reference points generated by group boundaries may only help tasks that require processing of the absolute serial positions such as serial recall and cued recall, but may not consistently facilitate relative order judgements, as the relevant reference point could be either be earlier, later, or across both probe items (for an example, see Introduction in Chapter 5). Our results suggest against this hypothesis, as we found grouping reduces error rate in both the serial recall and JOR paradigm, and we found support for a two-level positional coding mechanism in the JOR error rate results.

In sum, contrary to the data suggesting differential results between the serial recall and comparative judgement tests are caused by underlying differences between using absolute and relative order memory, the results of grouping studies suggest order memory tasks may be more similar than previous thought.

6.2.3 Comparing different behavioural measures

Our results in Chapter 4 and Chapter 5 directly contribute to the theoretical debate of whether response time results could be compared with error rate results, as we use the same grouping manipulations and collected data using both serial recall and the JOR paradigm. The grouping effects on the error rate measures are very similar between the JOR and serial recall paradigms. Grouping is found to decrease overall error rates for both paradigms, and there is evidence that within-group positions generate specific order errors in both paradigms. The finding of grouping effects on the JOR error rate and different results for error rates and response times suggests that comparative judgement tasks using only the response time measure may be inadequate to detect grouping effects, and one cannot assume response time and error rates reflect the same source of processing (Kahana & Loftus, 1999).

There is an additional reason for caution in comparing the response time measures between paradigms, as the response time measure typically used in serial recall is an inter-item response time, which is measured by the time differences between each recall response, whereas the response time measure used in JOR is the time between the onset of the probe and making a response. Thus, for serial recall, the retrieval process could be done before outputting the sequence and the inter-response time may reflect only output speed, not processing speed. This hypothesis is supported by an increase of initial response time as a function of list length, and relative constant inter-response time following the first items (Thomas et al., 2003) and a low correspondence between inter-response times and error rates.

The response time pattern generally supports a two phase recall strategy, that there is an assembly phase, where information is processed, followed by the ballistic output of assembled items, where the already recalled items are outputted by a motor response (Dennis, 2009). The results are also compatible with ACT-R (J. A. Anderson et al., 1998) by assuming first processing list-level information, followed by processing item-level information. In Chapter 4, we found slowed first retrieval, at both the beginning of the list and beginning of each group, generally supporting the two-phase explanation. However, the inter-response times are not constant after the first retrieval, suggesting further assumptions are required for modelling the inter-response time results. For the JOR task, the response time for making each judgement reflect the time to make decisions as well as output the decisions; thus, speed-accuracy tradeoffs should be found at each serial position. More research needs to be done to investigate the difference between response time and error rate measures in the JOR paradigm. We will further discuss possible mechanisms to account for both the response time and error rate patterns in the next section.

6.2.4 Grouping and backward serial recall

The effects of grouping have been studied predominantly using forward recall. Thomas et al. (2003) and Haberlandt et al. (2005) suggest in backward recall, participants use a scan-and-drop strategy. The scan-and-drop strategy could be described as participants repeatedly search from the beginning of the list towards the end of the list, the scanning process repeats when the target item is outputted and dropped from the list, until all items are recalled. Our results did not support the use of scan-and-drop strategy. One possible explanation of this discrepancy is that the use of scan-and-drop strategy is not viable for supraspan lists, and this strategy may only be found in subspan lists as demonstrated by Thomas et al. (2003) and Haberlandt et al. (2005). In the grouped lists, forward and backward recall error rate serial position curves are qualitatively similar when aligned by output position, which suggests the list could be outputted directly in backward order, and position codes maybe processed in either forward or backward direction. The inter-response time pattern for forward and backward recall of grouped lists is also consistent with the hypothesis that recall is dominated by output position. The inter-response time for backward recall is slower than forward recall, suggesting although position codes could be processed directly in backward order, they could be handled differently.

6.3 Implications for models of order memory

The JOR paradigm is in a unique position to bridge theoretical work from comparative judgements to serial recall. Our results suggest commonality of the two order memory tasks and existing memory models should be updated to account for the newly established behavioural findings. We discuss how current memory models could be adapted to explain the congruity effect, speed-accuracy tradeoffs, backward serial recall and grouping effects.

6.3.1 Congruity effect

The JOR congruity effect is a recent finding that has been overlooked by order memory modellers. To the author's knowledge, only two models were fitted for the JOR results prior to the findings of the congruity effect, that is Hacker's (1980) self-terminating search model and OSCAR (Brown et al., 2000), and both of those models are based on a search or matching process starting from the end of the list towards the beginning. In Chapter 2 (Liu et al., 2014) we found for subspan lists the congruity effect could support a search direction switch by implementing Hacker's model with a forward search direction for "earlier" instruction and a backward search direction for "later" instruction. However, the directionality switch does not account for the supraspan results, as the behavioural pattern is dominated by an overall recency effect. In Chapter 4 we found some support that congruity effect could be related to recall direction, as backward recall serial position curves are mirror images of forward recall when plotted against input position. However, it is unclear how serial recall and JOR could be related. Models that depend on direct inter-item associations (e.g., TODAM, Murdock, 1995) could also model the congruity effect of subspan JOR by assuming recall starts from either the beginning of the list for the "earlier" instruction or the end of the list for the "later" instruction. However, those models may have difficulty to account for the distance effect, as item access is dominated by list position. The congruity effect might be easier to implement in positional coding models, where items are not directly associated with each other, but are associated with a positional marker. The positional coding could be either temporal (e.g., Brown et al., 2000, 2007), relational (Henson, 1998) or absolute (e.g., Farrell, 2012b). In those models the congruity effect may be accounted for by adding an additional dimension of position, where the added dimension represents a bias towards either the beginning or end of the list, or by adding a factor that systematically distorts the temporal dimension.

In an unpublished analysis (see Chapter 2.7), we fitted results from Chapter 2 (Liu et al., 2014) to Scale Independent Memory Perception and Learning (SIMPLE) model (Brown et al., 2007) and found adding a bias gradient that systematical alters the temporal codes that selectively enhanced position discriminability either towards the beginning or end of the list could account for the congruity effect. This SIMPLE model is consistent with our findings that the congruity effect is found to generalize across wide range of factors, and accounts for a significant source of variance in addition to other known JOR benchmark effects. We suggest that for models that rely on positional coding, the congruity effect could be modelled by adding a factor that systematically changes the list-level position codes, and possibly by adding a separate bias dimension.

Models developed for the comparative judgement paradigm were not specific for order memory, but may shed some light on how to implement the congruity effect in memory models. The congruity effect may be a phenomenon associated with the comparison process, which in turn influences the relative order judgement. A prominent idea to account for the congruity effect in comparative judgements is the reference point theory (Holyoak, 1978). Reference points refer to points on a continuum that other values can take as a reference, and usually depends on the behavioural task's instruction (e.g., the question "Which number is larger than 5?" explicitly sets the reference point to be 5, and all other numbers would compare to 5.). The extreme members of a continuum are considered the default reference points when no specific reference point is provided. The model assumes each judgement is made by comparing both stimuli to a reference point. The ratio of differences between stimuli and reference point can account for the congruity effect. Other variants of reference point theory have different assumptions (e.g., Petrusic, 1992; Marks, 1972), but all focus on the same general concept. Although not obvious how this can be implemented, future modelling development could test the reference point concept as a core mechanism for congruity effect. The challenge of applying the reference point mechanism is that it is always unclear what position is used as a reference for making a relative order judgement. We will revisit this concept in model discussion of grouping effects.

6.3.2 Speed-accuracy tradeoffs

In the JOR paradigm, the response times and error rates may reflect a wide range of different processes. Factors underlying response times include the access speed of order memory, the

processing time for making judgement, as well as the execution of motor responses, whereas the error rate measure may reflect the quality of order memory, competitions between items or context cues, and guessing. Under most circumstances, it is reasonable to assume that when items are easier to access and are of good quality, both response time and error rate should be lower, and vice versa when items are harder to access and is of poor quality. This general assumption holds true for results in Chapter 2 (Liu et al., 2014); however, we found evidence of speed-accuracy tradeoffs for the English alphabet in Chapter 3 as well as for grouped lists in Chapter 5. In Chapter 5, we found grouping enhances error rates, at a cost of slower respond times. The overall results suggest that although response time and error rate generally go in the same direction, an extreme speed-accuracy tradeoff could flip the direction of response time and error rates. The different underlying mechanisms for response time and error rates may have different roles in the speed-accuracy tradeoffs.

Memory models developed from serial recall generally do not address speed-accuracy tradeoffs (e.g., SIMPLE Brown et al., 2007) or cannot account for extreme speed-accuracy tradeoffs because the model predict response times are derived or positively correlated with the accuracy measure (e.g., OSCAR Brown et al., 2000). Data from serial recall often has near ceiling accuracy levels, this ceiling accuracy effect may decouple response-time from error-rate measures. The ceiling accuracy effect challenges model assumptions that derive response times based on accuracy. When response times are not derived from error rates, a model could account for speed-accuracy tradeoffs with additional assumptions (e.g., ACT-R; J. A. Anderson et al., 1998). For the ACT-R model, slow inter-item response times are explained by additional processing of multiple layers of position codes. The lack of direct mechanism for speed-accuracy tradeoffs is predominantly caused by the serial recall paradigms rely on inter-item response time. The response time for the first word recall may reflect retrieval of the whole list, and inter-item response time following the first recall may simply reflect how fast a person can execute the motor output (Dennis, 2009; Thomas et al., 2003). According to ACT-R model, a long pause before outputting each group may reflect retrieval of the full chunk, and the increased response time after the initial pause may reflect the necessary time to output each item. For this reason, the inter-item response time and accuracy measure different mechanisms, and speed-accuracy tradeoffs may be irrelevant. For the JOR paradigm, both response time and error rate may reflect the retrieval of items, and it is possible a common mechanism could account for extreme speed-accuracy tradeoffs that

the two measures change in opposite directions.

Because the researchers who model comparative judgements typically only measure response time, the speed-accuracy tradeoffs are also overlooked by most models. Researchers using the comparative judgement paradigm typically use a random walk process to model both response time and error rates (Ratcliff, 1978; Birnbaum & Jou, 1990). The random walk process could predict speed-accuracy tradeoffs by manipulation of the initial onset point of the random walk as well as set different threshold boundaries, but could never predict both measures change in opposite direction. Petrusic (1992) first bring up this issue by showing the congruity effect could be found on the error rate measure, and proposed Slow and Fast Guess Theory (SFGT) to address both the response time and error rate pattern and their extreme tradeoffs. The SFGT model suggests information for making the correct judgement accumulates as time increases, and the rate of information accumulation is faster when the quality of information is good. This generally predicts faster response time is associated with lower error rates and slower response time is associated with higher error rates. However, participants may decide to guess when they are under time pressure (fast guess) or when the information quality is too low it takes very long time without reaching a decision (slow guess). An extreme speed-accuracy tradeoff pattern could be attributed to the information accumulation rate as well as the proportion of both fast and slow guess. The SFGT mechanism is compatible with positional coding theories, as we already know the rate of information accumulation is related to instruction, distance, serial position and grouping, where all four factors could be derived from hierarchical position codes. It remains a future challenge for implementing SFGT to figure out the how much evidence is required to make a decision and when to guess, as both thresholds and information accumulation affects response times and error rates. In sum, it is important to analyze both response time and error rate results at the same time to account for possible speed-accuracy tradeoffs. One can not draw any conclusion from just the response time measure, as the error rate measure may show opposite pattern under circumstances. e.g., grouping does facilitate error rates but hinders response times in Chapter 5. The SFGT model provides a possible explanation of the differential speed-accuracy tradeoffs without proposing a separate model mechanism for generating response times.

6.3.3 Grouping effect

In Chapter 4 we replicated previous results of grouping effects on serial recall, and found enhanced accuracy and faster response times for the grouped lists than ungrouped lists. We further show grouping selectively reduces adjacent transposition errors and the serial position curves for forward and backward serial recall are mirror images. In Chapter 5, we further showed the grouping effect on error rates in the JOR paradigm are very similar to the effects found in serial recall, and show support of two-level positional coding models. The sets of findings add further constraints to current memory models.

In the grouped serial recall literature, grouping is found to enhance accuracy at all serial positions, and the same effect have been found across different list materials and modalities (Henson, 1998; Hitch et al., 1996; Frankish, 1985, 1989; Maybery et al., 2002). However, we found grouping does not enhance accuracy at 4 of the 9 serial positions, at positions 1, 2, 4, and 5. The difference in grouping effects could be attributed to the total list presentation time. Majority of the studies induced grouping by adding additional temporal pauses, without controlling for the total duration (e.g., Frankish, 1985, 1989; Hitch et al., 1996; Henson, 1998), this allows more overall rehearsal time for the grouped lists than the ungrouped lists, and this may underly overall accuracy advantage of the grouped lists. In studies where the total presentation is matched between grouped and ungrouped lists, the grouping effects interact with serial positions, where the grouping advantage is reduced for the beginning and the end of the lists (Maybery et al., 2002). In Chapter 4 we matched the total presentation time for grouped and ungrouped lists, thus our results could directly test models based on the dimension of time, such as SIMPLE (Brown et al., 2007). Our results challenge SIMPLE because the group boundary produce enhanced discriminability for items presented either before or after the temporal pause, thus SIMPLE could lead one to predict higher accuracy for boundary items. This is not the case in our data, where grouping selectively enhance the end items of a group 1 and group 2. This pattern could also be found in Brown et al.'s (2007) model fit of the von Restorff effect, where the adjacent items of the distinctive items are not enhanced in the data, but are enhanced by SIMPLE. The finding against SIMPLE is inconsistent with the assumption that grouping enhances precision of item positions, as assumed by the two-level positional coding models. Future modelling work should address the selective enhancement of items by grouping.

Our finding of output order dominating the serial recall pattern for the grouped lists cannot be easily accounted for by a retrieval mechanism assuming forward directional processing of position codes and we found no evidence of the use of scan-and-drop strategy (Thomas et al., 2003) to recall grouped list backwards. The results suggest positional coding models depending on temporal code (e.g., SIMPLE Brown et al., 2007), ordinal code (e.g., ACT-R J. A. Anderson et al., 1998) or relative code (Henson, 1998) may need further development to account for retrieval strategies of groups in backward order. The positional coding models may need flexible retrieval mechanisms that can process position codes in backward order.

We are the first to demonstrate grouping effects of error rates are very similar between the serial recall and the JOR paradigm. Similar to the serial recall results, we found support for two-level position codes in Chapter 5. The finding of group effects differ from Jou's (2011) argument that the group structure is not useful for making relative order judgements. It is possible that the grouping effects is more sensitive to the error rate measure, where in comparative judgements the accuracy is typically trained to a criterion. We suggest a unified account of the grouping effects of order memory is possible, with challenges. For the JOR paradigm, both the response time and error rate patterns are important, as they show differential results, thus the candidate model need to account for the differential patterns from error rates and response times in both paradigms. In addition, "earlier" and "later" instruction for the JOR paradigm are not equivalent to "forward" and "backward" instruction in serial recall, as the JOR task is asking a relative order question where recall direction is not strictly enforced, whereas serial recall is asking participants to recall the list in order. Thus, it is important to understand the underlying mechanism of instruction and grouping instruction.

6.3.4 Future modelling directions

As mentioned earlier, both the JOR and serial recall results support the positional coding models. Thus, a possible solution for modelling JOR results is to adapt the positional coding models to the JOR paradigm, to account for the congruity effect, distance effect, serial position effects, grouping effect as well as the speed-accuracy tradeoffs in the JOR paradigm. We suggest the additional benchmark effects found in the JOR paradigm might be cause by a decision making process that is independent from order memory. SFGT theory developed for the comparative judgement paradigm is a candidate of a decision making model, that

already can account for the speed-accuracy tradeoffs found in the JOR paradigm. Thus, one possible solution is to combine positional coding concepts with the FSGT theory, so that response time and error rates could both be derived from hierarchical position codes.

6.4 Limitations

In this section we discuss the limitations of the current studies and how could those limitations be addressed.

First, the choice of the stimuli set may influence the overall findings. The nouns and consonant pairs we used for the study lists already have pre-existing semantic associations, which could interact with the temporal learning and judgements. The consonant lists could be perceived as a set letters that have a pre-existing meaning, and the two item probe could be perceived as a bi-gram that also has meaning. For example, for a four letter list DKNY is a pre-existing brand name. The probe BF and FB could be associated with "boyfriend" or "facebook". Because the set of the stimuli pool was limited and items were re-sampled, a part of the high error rates could be attributed to interferences. The pre-existing associations may also form chunks, for example MLNOP from the English alphabet might be perceived a chunk, and the chunk structure may interfere with how people make order judgements for probes within-chunk or between-chunk. To address this, we could include bi-gram frequency as a factor in our analysis. Ideally, to help disentangle whether the congruity effect is not a by-product of the pre-existing associations, we could use more abstract stimuli such as abstract pictures and colours.

Second, we could not rule out the possibility that the congruity effects found on subspan and supraspan lists are not generalization of the same effect. Although the form of instruction by serial position interaction is the same, it is possible that for subspan lists the congruity effect could be caused by self-terminating searches of opposite directions, and for supraspan lists the congruity effect has a different underlying cause. One way of testing whether the congruity effect is the same between subspan and supraspan lists is to enforce the use of self-terminating searches in supraspan lists and compare the congruity effect to a control list without self-terminating search enforcements. The congruity effect generated by self-terminating search should be systematically different from the congruity effect found on the control list. In addition, if the congruity effect for subspan and supraspan lists were different,

it is unclear what governs the strategy change from self-terminating search to direct access. We hypothesize the strategy might be changed depending on whether accuracy could be maintained at ceiling. A self-terminating search strategy is viable and easy to perform if the participants have perfect memory of the list. However, as the list gets longer, ceiling accuracy could not be achieved, and self-terminating search should take longer to perform as list length increases. It is also possible that the congruity effects of supraspan list is a result of a mix of strategies. For example, probes within the same group might be processed by self-terminating search rather than a direct access based strategy. Further analysis of the data from Chapter 3 and Chapter 5 may tell us more about whether there is evidence of a mixed strategy use. A direct test of mixed strategies may require experimental manipulation of the strategy used, thus it is important to find a reliable method of manipulating JOR strategies for future studies.

Finally, our understanding of speed-accuracy tradeoffs is still limited by our current dataset. No evidence of speed-accuracy tradeoffs were found in Chapter 2 (Liu et al., 2014), yet we found speed-accuracy trade-offs in Chapter 3 and Chapter 5. It is unclear why speed-accuracy tradeoff differently for the first and second half of the English alphabet. The data from Chapter 5 suggest speed-accuracy trade-offs could be related to grouping of the list, where grouping lowers the error rates at the cost of slower response times. Because of the limited number of empirical studies using error rate measures, and we have only tested one grouping structure and one list length in Chapter 5, we do not know whether the results can generalize to lists with different stimuli types, list lengths, and grouping structures. We are bridging from the JOR paradigm to the serial recall paradigm based on the error rate measure, thus it is especially important for us to understand the underlying factors and boundary conditions of JOR speed-accuracy tradeoffs. At this point, we could not answer whether grouping always slows response times for JOR if we grouped the list differently.

6.5 Future directions

For future research, it is important to distil the most relevant theoretical questions that we could not already answer right now. One direction we can pursue is to ask whether there is a common mechanism for the congruity effect across span. Regardless if we found a common mechanism, it is also important to model and explain what the underlying mechanism

is. Another direction is to explore the factors underlying speed-accuracy tradeoffs and its relation to other benchmark effects. The immediate question we could ask is whether speed-accuracy tradeoffs found on grouped lists could be replicated. The long term question is whether the wide range of benchmark effects on JOR and serial recall could be account for by one memory model.

To summarize, the research presented in this dissertation systematically established the congruity effect as a benchmark effect of JORs, and expanded our knowledge of how JORs, comparative judgements, and serial recall results are interrelated from the results of grouping studies. The converging evidence of from JOR data sets suggest a common mechanism underlying human order memory paradigms is possible, and JOR benchmark effects should be considered seriously for developing order memory theories.

Bibliography

Alloway, T. P., Gathercole, S. E., & Pickering, S. J. (2006). Verbal and visuo-spatial short-term and working memory in children: Are they separable? *Child Development*, *77*, 1608-1716.

Anderson, D. R., & Burnham, K. P. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research*, *33*, 261-304.

Anderson, J. A., Silverstein, J., Ritz, S. A., & Jones, R. S. (1998). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review*, *84*, 413-451.

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

Asch, S. E., & Ebenholtz, S. M. (1962). The principle of associative symmetry. *Proceedings of the American Philosophical Society*, *106*, 135-163.

Audley, R. J., & Wallis, C. P. (1964). Response instructions and the speed of relative judgement. I. some experiments on brightness discrimination. *British Journal of Psychology*, *55*, 59-73.

Baayen, R. H. (2007). LanguageR (R package on CRAN version 1.1) [Computer software and manual]. `http://cran.r-project.org/web/packages/languageR/index.html`.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390-412.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12-28.

Baker, R., Tehan, G., & Tehan, H. (2012). Word length and age influences on forward and backward immediate serial recall. *Memory & Cognition*, *40*, 40-51.

Banks, W. P. (1977). Encoding and processing of symbolic information in comparative judgments. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. 11, p. 101 - 159). Academic Press. doi: DOI:10.1016/S0079-7421(08)60476-4

Banks, W. P., & Flora, K. (1977). Semantic and perceptual processes in symbolic comparison. *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 278-290.

Banks, W. P., & Root, M. (1979). Semantic congruity effects in judgments of loudness. *Perception & Psychophysics*, *26*, 133-142.

Banks, W. P., White, H., & Mermelstein, R. (1980). Position effects in comparative judgments of serial order: List structure vs. differential strength. *Memory & Cognition*, *8*, 623-630.

Banks, W. P., White, H., Sturgill, W., & Mermelstein, R. (1983). Semantic congruity and expectancy in symbolic judgments. *Journal of Experimental Psychology: Human*

Perception and Performance, *9*(4), 560-582.

Bates, D. M. (2005). Fitting linear mixed models in R. *R News*, *5*, 27-30.

Bates, D. M., & Sarkar, D. (2007). lme4: Linear mixed-effects models using s4 classes (version 0.999375-39) [Computer software and manual]. `http://cran.r-project.org/web/packages/lme4/`.

Beaman, C. P. (2002). Inverting the modality effect in serial recall. *The Quarterly Journal of Experimental Psychology*, *55A*(2), 371-389.

Beaman, C. P., & Morton, J. (2000). The separate but related origins of the recency effect and the modality effect in free recall. *Cognition*, *77*, 59-65.

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.

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

Birnbaum, M. H., & Jou, J. (1990). A theory of comparative response times and "difference" judgments. *Cognitive Psychology*, 184–210.

Bower, G. H. (1971). Adaptation-level coding of stimuli and serial position effects. In M. H. Appley (Ed.), (p. 175-201). New York: Academic Press.

Brannon, E. M. (1997). *Chunking in memory: Toward an operational definition of a chunk.*

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

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

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel interference* (Second ed.). New York: Springer-Verlag.

Butters, M. A., Kaszniak, A. W., Glisky, E. L., Eslinger, P. J., & Schacter, D. L. (1994). Recency discrimination deficits in frontal lobe patients. *Neuropsychology*, *8*(3), 343-353.

Caplan, J. B. (2015). Order-memory and association-memory. *Canadian Journal of Experimental Psychology*, *69*(3), 221-232.

Caplan, J. B., Glabolt, M. G., & McIntosh, A. R. (2006). Linking associative and serial list memory: Pairs versus triples. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(6), 1244-1265.

Cattell, J. M. (1902). The time of perception as a measure of differences in intensity. *Philosophische Studien*, *19*, 63-68.

Cech, C. (1995). Congruity and the expectancy hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 1275-1288.

Cech, C., & Shoben, E. J. (2001). Categorization process in mental comparisons. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(3), 800-816.

Chan, M., Ross, B., Earle, G., & Caplan, J. B. (2009). Precise instructions determine participants' memory search strategy in judgments of relative order in short lists. *Psychonomic Bulletin & Review*, *16*, 945-951.

Conrad, R. (1965). Order error in immediate recall of sequences. *Journal of Verbal Learning and Verbal Behaviour*, *4*, 161-169.

Crowder, R. G. (1982). The demise of short-term memory. *Acta Psychologica*, *50*, 291-323.

Dennis, S. (2009). Can a chaining model account for serial recall? In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the thirty-first annual meeting of the cognitive science society* (p. 2813-2818). Austin, TX: Cognitive Science Society.

Duncan, E. M., & McFarland, C. E. (1980). Isolating the effects of symbolic distance and semantic congruity in comparative judgements: An additive-factor analysis. *Memory & Cognition*, *8*, 612-622.

Ellis, S. H. (1972). Interaction of encoding and retrieval in relative age judgments: An extension of the "crossover" effect. *Journal of Experimental Psychology*, *94*, 291-294.

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*(2), 324-347.

Farrell, S. (2012a). Multiple roles for time in short-term memory: Evidence from serial recall of order and timing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 128-145.

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

Farrell, S., & Lelièvre, A. (2012). The dynamics of access to groups in working memory. *Journal of Experimental Psychology*, *38*(6), 1659-1674.

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

Flexser, J., & Bower, G. H. (1974). How frequency affects recency judgments: A model for recency discrimination. *Journal of Experimental Psychology*, *103*(4), 706-716.

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (Second ed.). Thousand Oaks CA: Sage. Retrieved from `http://socserv.socsci.mcmaster.ca/jfox/Books/Companion`

Fozard, J. L. (1970). Apparent recency of unrelated pictures and nouns presented in the same sequence. *Journal of Experimental Psychology: Human Learning and Memory*, *86*(2), 137-143.

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

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

Fuhrman, R. W., & Wyer, J. R. S. (1988). Event memory: Temporal-order judgments of personal life experiences. *Journal of Personality and Social Psychology*, *54*(3), 365-384.

Gelinas, C. S., & Desrochers, A. (1993). Positive and negative instructions in symbolic paired comparisons with the months of the year. *Psychological Research*, *55*, 40-51.

Geller, A. S., Schleifer, I. K., Sederberg, P. B., Jacobs, J., & Kahana, M. J. (2007). Pyepl: A cross-platform experiment-programming library. *Behavior Research Methods*, *39*(4), 950-958.

Gerton, B. K., Brown, T. T., Meyer-Lindenberg, A., Kohn, P., & Holt, J. (2004). Shared and distinct neurophysiological components of the digits forward and backward tasks

as revealed by functional neuroimaging. *Neuropsychologia*, *42*, 1781-1787.

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.

Grenzebach, A. P., & McDonald, J. E. (1992). Alphabetic sequence decisions for letter pairs with separation of one to three letters. *Journal of Experimental Psychology*, *18*(4), 865-872.

Guérard, K., Saint-Aubin, J., Burns, S. C., & Chamberland, C. (2012). Revisiting backward recall and benchmark memory effect: A reply to bireta et al. *Memory & Cognition*, *40*, 338-407.

Haberlandt, K., Lawrence, H., Krohn, T., Bower, K., & Thomas, J. G. (2005). Pauses and durations exhibit a serial position effect. *Psychonomic Bulletin & Review*, *12*, 152-158.

Hacker, M. J. (1980). Speed and accuracy of recency judgements for events in sort-term memory. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(6), 651-675.

Henson, R. N. A. (1998). Short-term memory for serial order: the start-end model. *Cognitive Psychology*, *36*(2), 73–137.

Henson, R. N. A. (1999). Positional information in short-term memory: Relative or absolute? *Memory & Cognition*, *27*(5), 915–927.

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

Hinrichs, J. V. (1970). A two-process memory-strength theory for judgment of recency. *Psychological Review*, *77*, 223-233.

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

Hockley, W. (1984). Analysis of response time distribution in the study of cognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 598-615.

Holyoak, K. J. (1977). The form of analog size information in memory. *Cognitive Psychology*, *9*, 31-51.

Holyoak, K. J. (1978). Comparative judgements with numerical reference points. *Cognitive Psychology*, *10*, 203-243.

Holyoak, K. J., & Mah, W. A. (1981). Semantic congruity in symbolic comparisons: evidence against an expectancy hypothesis. *Memory & Cognition*, *9*, 197-204.

Holyoak, K. J., & Patterson, K. K. (1981). A positional discriminability model of linear order judgements. *Journal of Experimental Psychology: Human Perception and Performance*, *7*, 1283-1302.

Howard, M. W., & Kahana, M. J. (1999). Contextual variability and serial position effects in free recall. *Journal of Experimental Psychology*, *25*(4), 923-941.

Howard, R. W. (1980). Category use in abstract mental comparisons. *Quarterly Journal of Experimental Psychology*, *32*, 625-633.

Hulme, C., Roodenrys, S., Schweickert, R., Brown, G. D. A., Martin, S., & 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 and 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 Reviews*, *140*(2), 339-373.

Hurst, W., & Volpe, B. T. (1982). Temporal order judgements with amnesia. *Brain and Cognition*, *1*, 294-306.

Jamieson, D. G., & Petrusic, W. M. (1975). Relational judgements with remembered stimuli. *Perception & Psychophysics*, *18*, 373-378.

Jou, J. (1997). Why is the alphabetical middle letter in a multiletter array so hard to determine? memory processes in linear-order information processing. *Journal of Experimental Psychology: Human Perception and Performances*, *23*(6), 1743-1763.

Jou, J. (2003). Multiple number and letter comparison: Directionality and accessibility in numeric and alphabetic memories. *The American Journal of Psychology*, *116*, 543-579.

Jou, J. (2005). Memory retrieval tasks determine the serial position curves of linear order with categorical structures. *The American Journal of Psychology*, *118*(4), 525-565.

Jou, J. (2010). The serial position, distance, and congruity effects of reference point setting in comparative judgments. *The American Journal of Psychology*, *123*(2), 127-136.

Jou, J. (2011). Two paradigms of measuring serial-order memory: two different patterns of serial position functions. *Psychological Research*, *75*, 202-213.

Jou, J., & Aldridge, J. W. (1999). Memory representation of alphabetic position and interval information. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *25*, 680-701.

Kahana, M. J. (2002). Associative symmetry and memory theory. *Memory & Cognition*, *30*, 823-840.

Kahana, M. J. (2012). *Foundations of human memory.* New York, NY: Oxford University Press.

Kahana, M. J., & Caplan, J. B. (2002). Associative asymmetry in probed recall of serial lists. *Memory & Cognition*, *30*(6), 841-849.

Kahana, M. J., & Loftus, G. (1999). Response time versus accuracy in human memory. In R. J. Sternberg (Ed.), *The nature of cognition* (p. 322-384). Cambridge, MA: MIT Press.

Klahr, D., Chase, W. G., & Lovelace, E. (1983). Structure and process in alphabetic retrieval. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *9*, 462-477.

Klein, K. A., Shiffrin, R. M., & Criss, A. H. (2007). Putting context in context. In J. S. Nairne (Ed.), *The foundations of remembering: Essays in honor of Henry L. Roediger III.* Psychology Press.

Kleinfeld, D. (1986). Sequential state generation by model neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, *83*, 9469-9473.

Köhler, W. (1947). *Gestalt psychology.* New York: Liveright.

Kosslyn, S. M., Murphy, G. L., Bemesderfer, M. E., & Feinstein, K. J. (1977). Category and continuum in mental comparisons. *Journal of Experimental Psychology: General*, *106*, 341-375.

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), (p. 112-131). New York: Wiley.

Lee, C. L., & Estes, W. K. (1977). Order and position in primary memory for letter strings. *Journal of Verbal Learning And Verbal Behavior*, *16*(4), 395-418.

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

Leth-Steensen, C., & Marley, A. A. J. (2000). A model of response time effects in symbolic comparison. *Psychological Review*, *107*(1), 62-100.

Lewandowsky, S., & Farrell, S. (2008). Phonological similarity in serial recall: Constraints on theories of memory. *Journal of Memory and Languange*, *58*, 429-448.

Lewandowsky, S., & Murdock, B. B. (1989). Memory for serial order. *Psychological Review*, *96*, 25-57.

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

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

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

Link, S. W. (1990). Modelling imageless throught: The relative judgment theory of numerical comparisons. *Journal of Mathmatical Psychology*, *34*, 2-41.

Liu, Y. S., Chan, M., & Caplan, J. B. (2014). Generality of a congruity effect in judgements of relative order. *Memory & Cognition*, *42*(7), 1086-1105.

Lockhart, R. S. (1969). Recency discrimination predicted from absolute lag judgements. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 42-44.

Lovelace, E. A., & Snodgrass, R. D. (1971). Decision times for alphabetic order of letter pairs. *Journal of Experimental Psychology*, *88*(2), 258-264.

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

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

Maki, R. H. (1981). Categorization and distance effects with spatial linear orders. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *7*, 15-32.

Maki, R. H. (1982). Why do categorization effects occur in comparative judgment tasks? *Memory & Cognition*, *10*(3), 252-264.

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

Marks, D. F. (1972). Relative judgement: a phenomenon and theory. *Perception & Psychophysics*, *11*, 156-160.

Marshuetz, C. (2005). Order information in working memory: An integrative review of evidence from brain and behaviour. *Psychological Bulletin*, *131*(3), 323-339.

Martin, E., & Noreen, D. L. (1974). Serial learning: Identification of subjective sequences. *Cognitive Psychology*, *6*, 421-435.

Masin, S. C. (1995). Probabilistic inferences, discrimination, and stimulus interference in comparative judgement. *Psychological Research*, *58*, 10-18.

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

McElree, B. (1996). Accessing short!term memory with semantic and phonological information: A time!course analysis. *Memory & Cognition*, *24*, 173-187.

McElree, B. (2006). Accessing recent events. In B. H. Ross (Ed.), (Vol. 46, p. 155 - 200). Academic Press.

McElree, B., & Dosher, B. A. (1993). Serial retrieval processes in the recovery of order information. *Journal of Experimental Psychology: General*, *122*(3), 291-315.

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

Milner, B. (1971). Interhemispheric differences in the localization of psychological processes in man. *British Medical Bulletin*, *27*, 272-277.

Motulsky, H., & Christopoulos, A. (2004). *Fitting models to biological data using linear and non-linear regression. a practical guide to curve fitting.* Oxford,UK: Academic Press.

Moyer, R. S., & Bayer, R. H. (1976). Mental comparison and the symbolic distance effect. *Cognitive Psychology*, *8*, 228-246.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature*, *215*, 1519-1520.

Murdock, B., Smith, D., & Bai, J. (2001). Judgments of frequency and recency in a distributed memory model. *Journal of Mathematical Psychology*, *45*(4), 564 - 602.

Murdock, B. B. (1974). *Human memory: Theory and data.* Potomac, MD: Lawrence Erlbaum.

Murdock, B. B. (1995). Developing todam: Three models for serial-order information. *Memory & Cognition*, *23*, 631-645.

Muter, P. (1979). Response latencies in discriminations of recency. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(2), 160-169.

Nairne, J. S. (2002). Remembering over the short-term: The case against the standard model. *Annual Review of Psychology*, *53*, 53-81.

Naveh-Benjamin, M. (1990). Coding of temporal order information: An automatic process? *Journal of Experimental Psychology: Learning, Memory and Cognition*, *16*(1), 117-126.

Neath, I., & Crowder, R. G. (1996). Distinctiveness and very short-term serial position effects. *Memory*, *4*(3), 225-242.

Neath, I., & Surprenant, A. (2003). *Human memory: An introduction to research, data, and theory.* Thomson/Wadsworth. Retrieved from `http://books.google.ca/books?id=xZ9WAAAAYAAJ`

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, *7*(4), 308-313.

Ng, H. L., & Maybery, M. T. (2002). Grouping in short-term verbal memory: Is position coded temporally? *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *55*, 391-424.

Ng, H. L., & Maybery, M. T. (2005). Grouping in short-term memory: Do oscillators code the positions of items? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 175-181.

Nipher, F. E. (1878). On the distribution of errors in numbers written from memory. *Transactions of the Academy of Sciences of St. Louis*, *3*, CCX-CCXI.

Paivio, A. (1975). Perception comparisons through the mind's eye. *Memory & Cognition*, *3*, 635-647.

Parmentier, F. B. R., King, S., & Dennis, I. (2006). Local temporal distinctiveness does not benifit auditory verbal and spatial serial recall. *Psychonomic Bulletin & Review*, *13*(3), 458-465.

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, and Cognition*, *34*, 1349-1355.

Petrusic, W. M. (1992). Semantic congruity effects and theories of the comparison process. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 962-986.

Petrusic, W. M., & Baranski, J. V. (1989). Semantic congruity effects in perceptual comparisons. *Perception & Psychophysics*, *45*, 439-452.

Petrusic, W. M., Shaki, S., & Leth-Steensen, G. (2008). Remembered instructions with symbolic and perceptual comparisons. *Perception & Psychophysics*, *70*, 179-189.

Plant, R. R., & Turner, G. (2009). Millisecond precision psychological research in a world of commodity computers: New hardware, new problems? *Behaviour Research Methods*, *41*(3), 598-614.

Pliske, R. M., & Smith, K. H. (1979). Semantic categorization in a linear order problem. *Memory & Cognition*, *7*(4), 297-302.

Pohl, R. F. (1990). Position effects in chunked linear orders. *Psychological Research*, *52*, 68-75.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59-108.

Reynolds, C. R. (1997). Forward and backward memory span should not be combined for clinical analysis. *Archives of Clinical Neuropsychology*, *12*, 29-40.

Richardson, J. T. (2007). Measures of short-term memory: A historical review. *Cortex*, *43*, 635-650.

Riedel, U., Kühn, R., & van Hemmen, J. L. (1988). Temporal sequences and chaos in neural nets. *Physical Review A*, *38*, 1105-1108.

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

Rosen, V. M., & Engle, R. W. (1997). Forward and backward serial recall. *Intelligence*, *25*(1), 37-47.

Ryan, J. (1969a). Grouping and short-term memory: Different means and patterns of grouping. *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.

Sailor, K. M., & Shoben, E. J. (1993). Effects of category membership on comparative judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*,

19, 1321-1327.

Scharroo, J., Leeuwenberg, E., Stalmeier, P. F. M., & Vos, P. G. (1994). Alphabetic search: Comment on klahr, chase, and lovelace(1983). *Journal of Experimental Psychology: Learning, Memory and Cognition*, *20*, 236-244.

Schweickart, O., & Brown, N. R. (2013). Magnitude comparison extended: How lack of knowledge informs comparative judgments under uncertainty. *Journal of Experimental Psychology: General*.

Shoben, E. J., Cech, C., Schwanenflugel, P. J., & Sailor, K. M. (1989). Serial position effects in comparative judgments. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(2), 273-286.

Shoben, E. J., & Wilson, T. (1998). Categorization in judgments of relative magnitude. *Journal of Memory and Languange*, *38*(1), 94-111.

Skowronski, J. J., Ritchie, D. T., Walker, W. R., Sedikides, C., Bethencourt, L. A., & Martin, A. L. (2007). Ordering our world: The quest for traces of temporal organization in autobiographical memory. *Journal of Experimental Social Psychology*, *43*, 850-856.

Skowronski, J. J., Walker, W. R., & Betz, A. L. (2003). Ordering our world: An examination of time in autobiographical memory. *Memory*, *11*(3), 247-260.

Sompolinsky, H., & Kanter, I. (1986). Temporal association in asymmetric neural networks. *Physical Review Letters*, *57*, 2861-2864.

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

Sternberg, S. (1975). Memory scanning: new findings and current controversies. *Quarterly Journal of Experimental Psychology*, *27*, 1-32.

Surprenant, A. M., Bireta, T. J., Brown, M. A., Jalbert, A., Tehan, G., & Neath, I. (2011). Backward recall and the word length effect. *The American Journal of Psychology*, *124*, 75-86.

Terrace, H. (2001). Chunking and serially organized behavior in pigeons, monkeys and humans. In R. Cook (Ed.), . Medford: MA: Comparative Cognition Press.

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

Tremblay, A. (2013). LMERConvenienceFunctions: a suite of functions to back-fit fixed effects and forward-fit random effects, as well as other miscellaneous functions (version 2.5) [Computer software and manual]. `http://cran.r-project.org/web/packages/LMERConvenienceFunctions/index.html`.

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*(5), 1207-1241.

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

Wicklgren, W. A. (1966). Associative instructions in short-term recall. *Journal of Experimental Psychology*, *72*, 853-858.

Wilson, M. D. (1988). The MRC psycholinguistic database: Machine readable dictionary, version 2. *Behavioral Research Methods*, *20*, 6-11.

Wolff, P. (1966). Trace quality in the temporal ordering of events. *Perceptual and Motor Skills*, *22*(1), 283-286.

Woocher, F. D., Glass, A. L., & Holyoak, K. J. (1978). Positional discriminability in linear orderings. *Memory & Cognition*, *6*, 165-173.

Wyer, R. S., Jr., Shoben, E. J., Fuhrman, R. W., & Bodenhausen, G. V. (1985). Event memory: The temporal organization of social action sequences. *Journal of Personality and Social Psychology*, *49*(4), 857-877.

Yntema, D. B., & Trask, F. P. (1963). Recall as a search process. *Journal of Verbal Learning and Verbal Behavior*, *2*(1), 65-74.

Zuur, A. F., Leno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R.* New York: Springer.