Generality of a congruity effect in judgements of relative order

Yang S. Liu

Department of Psychology, University of Alberta, Edmonton, Alberta

Michelle Chan

Department of Psychology, University of Alberta, Edmonton, Alberta

Jeremy B. Caplan

Department of Psychology and Centre for Neuroscience, University of Alberta, Edmonton, Alberta

We would like to thank Dr. Harald Baayen for statistics consultation and Christopher Madan for feedback on an earlier draft of the manuscript. We would also like to thank the Natural Sciences and Engineering Research Council (NSERC) of Canada and the Alberta Ingenuity Fund for funding support. Correspondence concerning this article should be addressed to Y. S. Liu, Department of Psychology, University of Alberta, Edmonton, AB, T6G 2E9 Canada (Tel: 780-492-7872, e-mail: ly6@ualberta.ca)

Abstract

The judgement of relative order (JOR) procedure is used to investigate serial-order memory. Measuring response times, the wording of the instruction (whether the earlier or the later item is designated as the target) reversed the direction of search in sub-span lists (Chan et al., 2009). If a similar congruity effect applied to above-span lists, and furthermore, with error rate as the measure, this could suggest how to model order-memory across scales. Participants performed JORs on lists of nouns (Experiment 1: list lengths = 4, 6, 8, 10) or consonants (Experiment 2: list lengths = 4, 8). In addition to the usual distance, primacy and recency effects, instruction interacted with serial position of the later probe in both experiments, not only in response time, but also in error rate, suggesting that availability, not just accessibility, is affected by instruction. The congruity effect challenges current memory models. We fitted Hacker's (1980) self-terminating search model and found a switch in search direction could explain the congruity effect for short lists but not longer lists. This suggests that JORs may need to be understood via direct-access models, adapted to produce a congruity effect, or a mix of mechanisms.

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 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 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 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 2d). The effect of instruction can be most clearly visualized if we plot the difference between the "earlier" and "later" instruction data (Figure 2e).

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 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 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

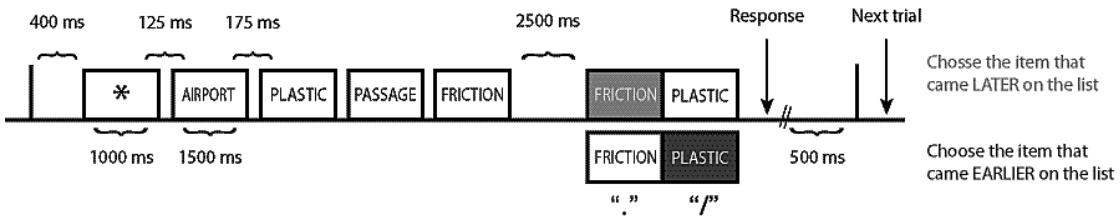

Figure 1. Time course of one example experimental trial in Experiment 1 (list length=4 nouns) with both instructions. At test, two nouns from the list are presented in random order, and the participant is asked to respond to the probe stimulus that occurred earlier ("earlier" instruction) or later ("later" 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.

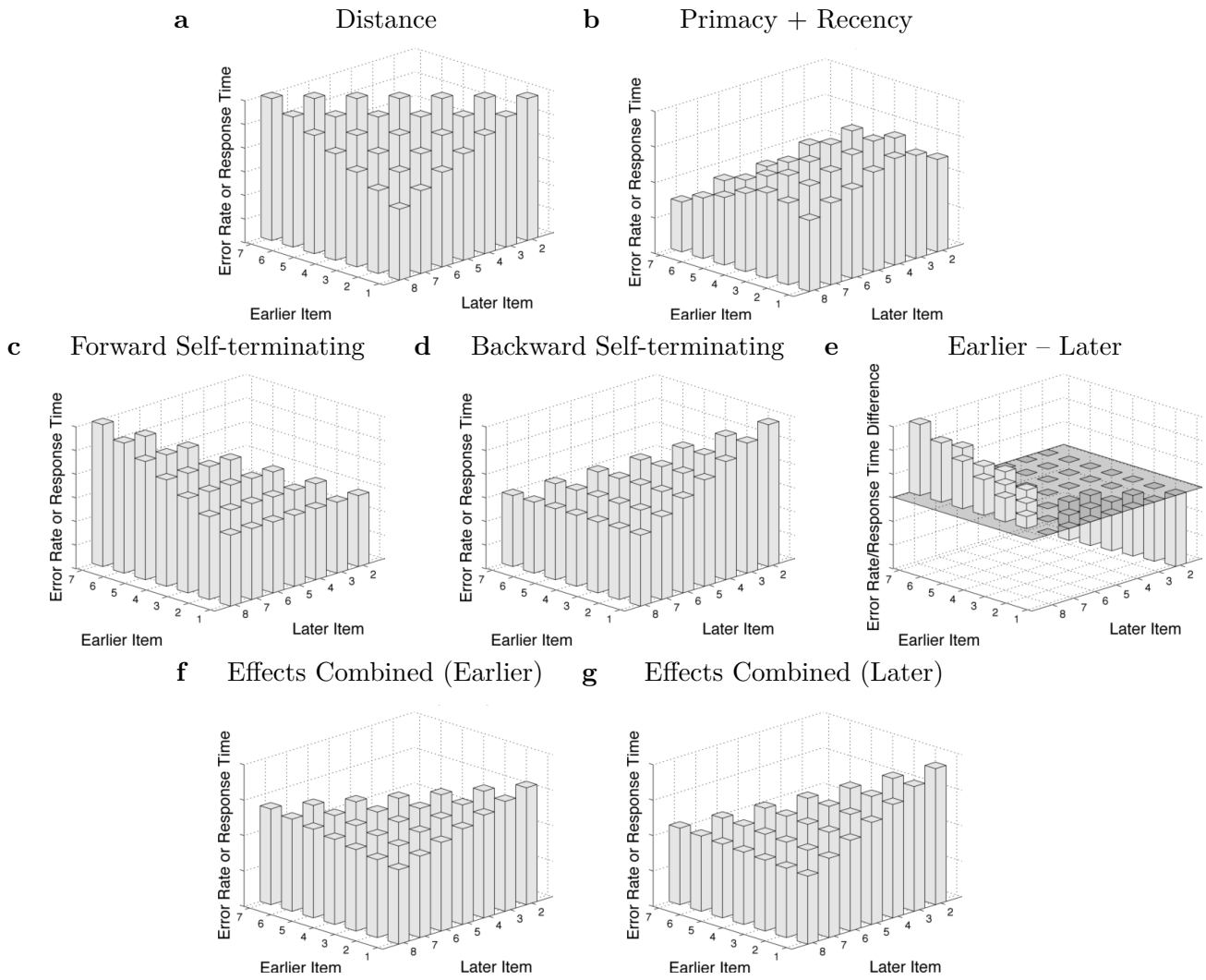

Figure 2. Schematic depictions of hypothesized serial position effects. The dependent measure (error rate or response time) is plotted as a function of both the earlier probe-item's serial position ("Earlier Item") and later probe-item's serial position ("Later Item"). **a,** Serial position effects 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 the "earlier" instruction (Chan et al., 2009). **d,** Serial position effects for backward, self-terminating search, as was found in sub-span lists using the "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 the "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 the "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.

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, Neath, & Chater, 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 the "earlier" instruction data and thus account for the congruity effect. Thus, the "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, Milner, & Haberlandt, 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 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 the "earlier" instruction, and backward, self-terminating search with the "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.

Experiment 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 the "earlier" or "later" group in alternating testing order. One participant in the "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 all capital letters. 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 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 the "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 the "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). 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 estimates random effects first, followed by fixed effects. In the results tables, the "Estimate" column reports the corresponding regression coefficients, along with their standard errors. For the purposes of reporting the LME results, the Intact condition and the "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).

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 3 and 4. We isolated the congruity effect by plotting the difference between the "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 1 and 2. To better visualize the pattern of serial-position effects, the overall mean was

		Estimate (SE)
Main effects		
	Intercept	-2.989 (0.287)*
	Intact/Reverse	0.5316 (0.090)*
	Later Probe Serial Position (Quadratic)	-51.39 (5.701)*
	Instruction	0.609 (0.393)
	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)	-79.77 (8.020)*
Interactions		
	Intact/Reverse × Instruction	-1.321 (0.132)*
	Trial × Session	0.122 (0.032)*
	Instruction × Later Probe Serial Position (Linear)	-35.44 (6.873)*
	Distance × Later Probe Serial Position (Linear)	26.47 (5.303)*
	LL × Later Probe Serial Position (Linear)	42.37 (5.999)*

Table 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 * - $p < 0.05$.*

removed to correct for the mean difference between the "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 3) showed a distance effect (Figure 2a), supported by a significant main effect of Distance, and bow-shaped serial position effect involving both primacy and recency (Figure 2b), supported by significant quadratic component of the Later Probe Serial Position in the best-fitting LME model (Table 1). The "later" instruction (Figure 3, middle column) broadly resembled the "earlier" instruction (Figure 3, left-hand column) except that the recency effect was more pronounced for the "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 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 1). This interaction was due to the "earlier" instruction producing better performance at earlier serial positions, and the "later" instruction producing better performance at later serial positions, in line

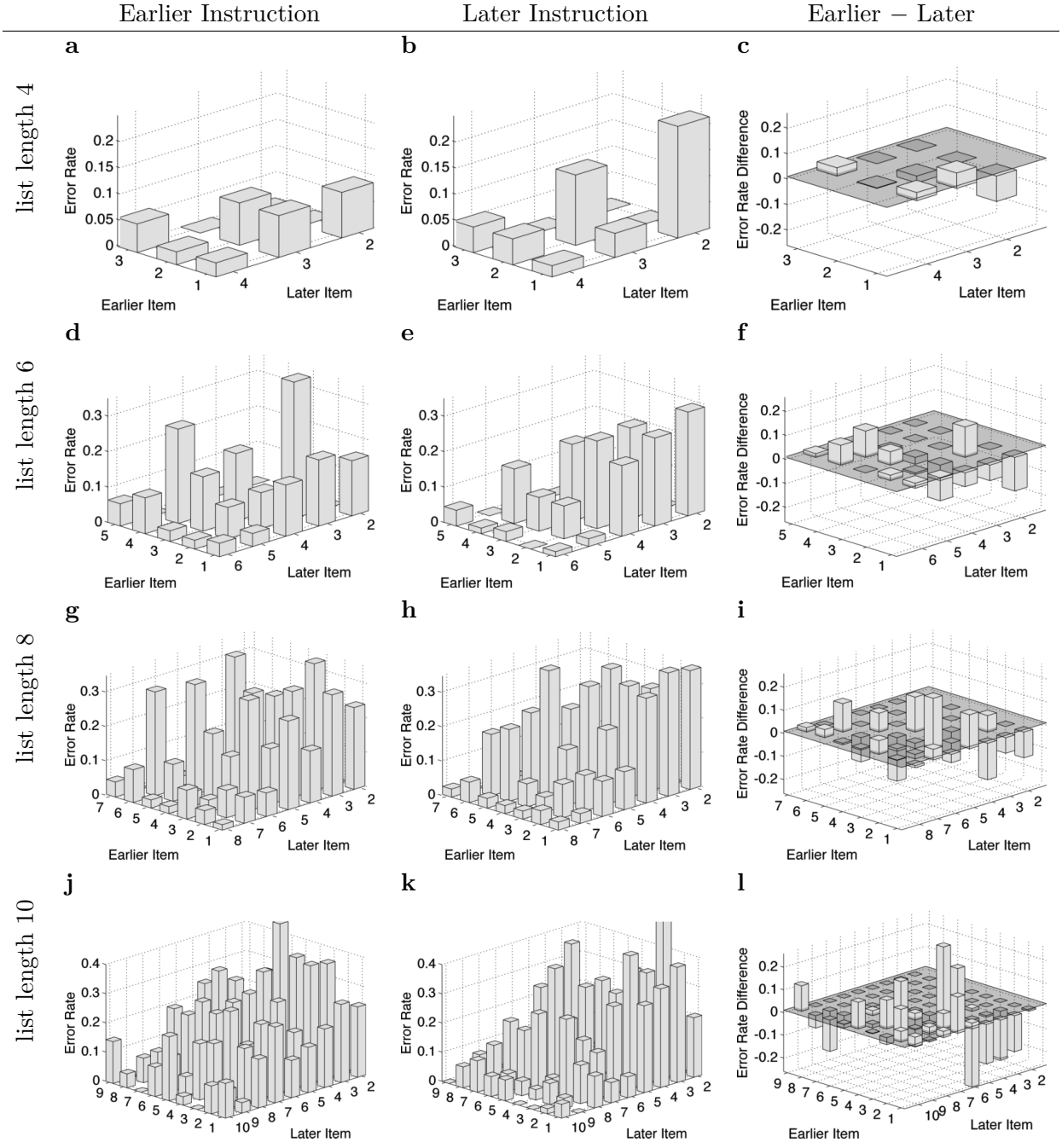

Figure 3. Error rate (Experiment 1) as a function of both probe items' serial position (earlier item and later item, respectively) broken down by list length in rows, and instruction ("earlier", "later" and the difference, "earlier"−"later", corrected for mean error rate) in columns.

with our predicted congruity effect (Figure 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.

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 the "earlier" instruction and worse for the "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 4) had significant distance and bow-shaped serial position effects (Figure 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).

Turning to the congruity effect, as with error rate, the difference bar graph (Figure 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). Again, in line with our predicted congruity effect (Figure 2e), the "earlier" instruction produced better performance at earlier serial positions, and vice versa for the "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 the "earlier" instruction than for the

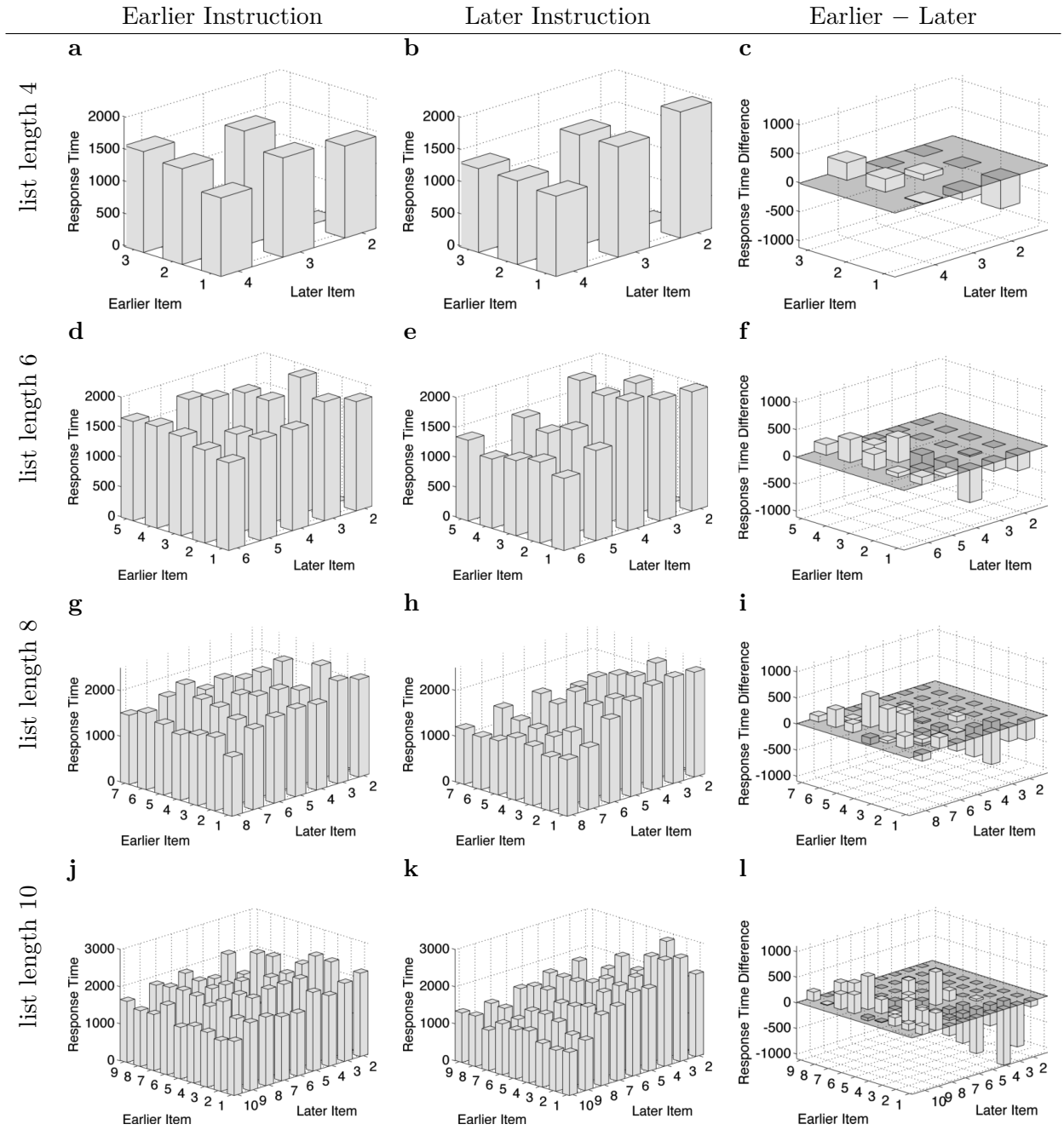

Figure 4. Response time (Experiment 1) as a function of both probe items' serial position (earlier item and later item, respectively) broken down by list length in rows, and instruction ("earlier", "later" and the difference, "earlier"–"later", corrected for mean response time) in columns.

		Estimate (SE)
Main effects		
	Intercept	7.20 (0.07)*
	LL	0.28 (0.01)*
	Instruction	0.075 (0.10)
	Intact/Reverse	0.038 (0.01)
	Trial	-0.015 (0.01)*
	Distance	-0.125 (0.01)*
	Session	-0.141 (0.01)*
	Later Probe Serial Position (Linear)	-17.5 (1.5)
	Later Probe Serial Position (Quadratic)	-18.1 (1.3)*
Interactions		
	LL × Instruction	0.056 (0.01)*
	LL × Distance	-0.015 (0.01)
	LL × Session	0.030 (0.00)*
	LL × Later Probe Serial Position (Linear)	10.7 (1.2)*
	Instruction × Intact/Reverse	-0.081(0.02)*
	Instruction × Trial	–0.032 (0.00)*
	Instruction × Distance	0.089 (0.02)*
	Instruction × Session	-0.030 (0.01)*
	Instruction × Later Probe Serial Position (Linear)	-13.7 (1.7)*
	Trial × Session	0.019 (0.00)*
	Distance × Later Probe Serial Position (Linear)	7.67 (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 (1.25)*
	Instruction × Distance × Later Probe Serial Position (Linear)	-6.13 (1.2)*
	LL × Instruction × Later Probe Serial Position (Quadratic)	-6.29 (1.0)*

Table 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 * - $p < 0.05$.*

"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 component of Later Probe Serial Position slope for the "later" Instructions and increasing quadratic component of Later Probe Serial Position slope for the "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 the "later" instruction improved more with practice than the in "earlier" instruction. The "later" instruction also had a smaller distance effect than the "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 the "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.

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.

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 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 capital letters. 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 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

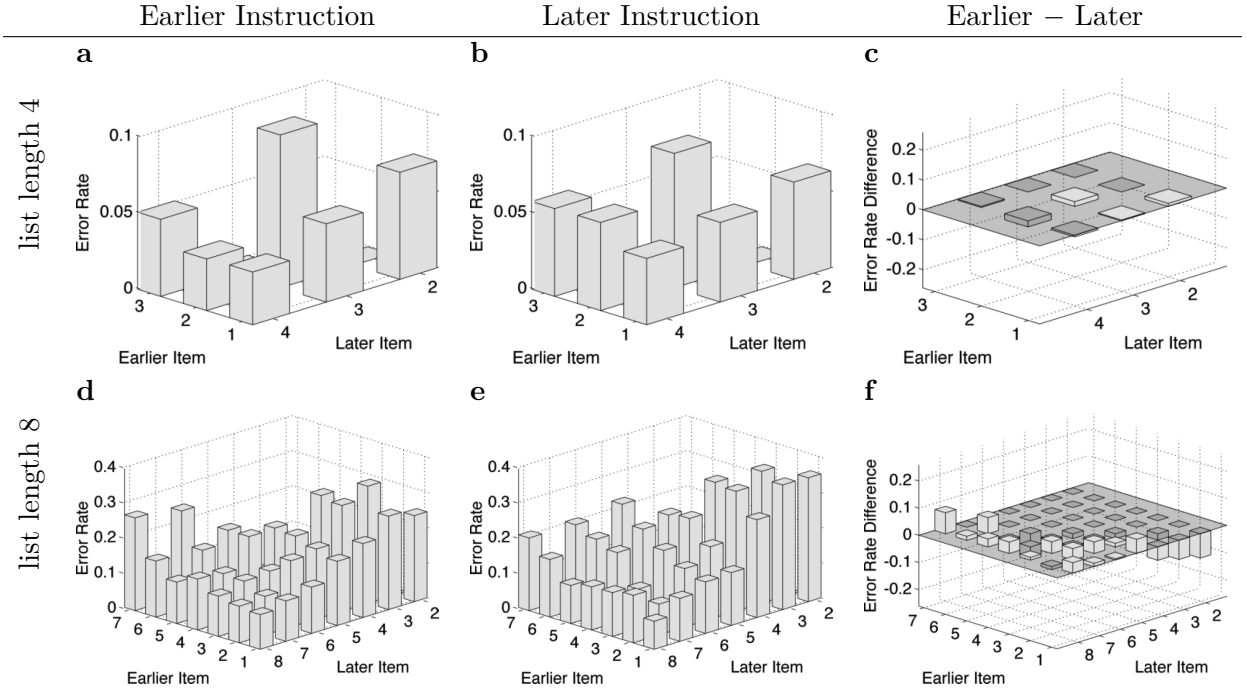

Figure 5. Error rate (Experiment 2) as a function of both probe items' serial position (earlier item and later item, respectively) broken down by list length in rows, and instruction ("earlier", "later" and the difference, "earlier"–"later", corrected for mean error rate) in columns.

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.

Results and Discussion

Error rates. First, because performance was near ceiling, we could not analyze error rates at list length 4 (Figure 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.

The list-length-8 data (Figure 5, bottom row) showed a congruity effect consistent with the

	BIC	AIC	Log-likelihood	Degrees of freedom
Best BIC model + Congruity effect	10119	19048	-9515.0	9
Best BIC model	10120	19057	-9520.6	8
Model difference ($\chi^2 = 11,\ p < 0.05$)	-1	-9	5.6	1

Table 3
Model comparison of best BIC model to best BIC 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.

		Estimate (SE)
Main effects		
	Intercept	-1.507 (0.06)*
	Intact/Reverse	0.405 (0.05)*
	Later Probe Serial Position (Linear)	-40.92 (6.18)*
	Instruction	0.735 (0.09)*
	Distance	-0.266 (0.02) *
	Trial	-0.134 (0.03)*
Interactions		
	Intact/Reverse × Instruction	-1.047 (0.07)*
	Instruction × Later Probe Serial Position (Linear)	-27.33 (8.20)*

Table 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 * - $p < 0.05$.*

pattern observed in experiment 1 (Figure 2e), with the "earlier" instruction resulting in more errors than the "later" instruction as Later Probe Serial Position increased, supported by a significant Instruction × Later Probe Serial Position (linear component) interaction in the best-fitting LME model (Table 4). For this LME model-selection, based on the BIC values, we cannot differentiate the lowest BIC model 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 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.

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

Table 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).

One can observe an overall recency effect at both list lengths (Figure 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 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. Probes in the "earlier" instruction were better judged than in the "later" instruction. This is despite more poor performers having been excluded for the "later" instruction (Table 5); thus, this indicates an overall advantage of the "earlier" instruction over the "later" instruction. Replicating experiment 1, the Intact/Reverse $\times$ Instruction congruity effect was also significant; intact probes were judged better for the "earlier" instruction and worse for the "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 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, the "earlier" instruction pattern resembled a distance effect with an overall primacy and recency effect (Figure 2f). The "later" instruction resembled a backward self-terminating pattern combined with distance, primacy and recency effects (Figure 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 6) showed a congruity effect (Figure 2e). The congruity effect is supported in the best-fitting model by a significant interaction of Instruction $\times$ Later Probe Serial Position (linear

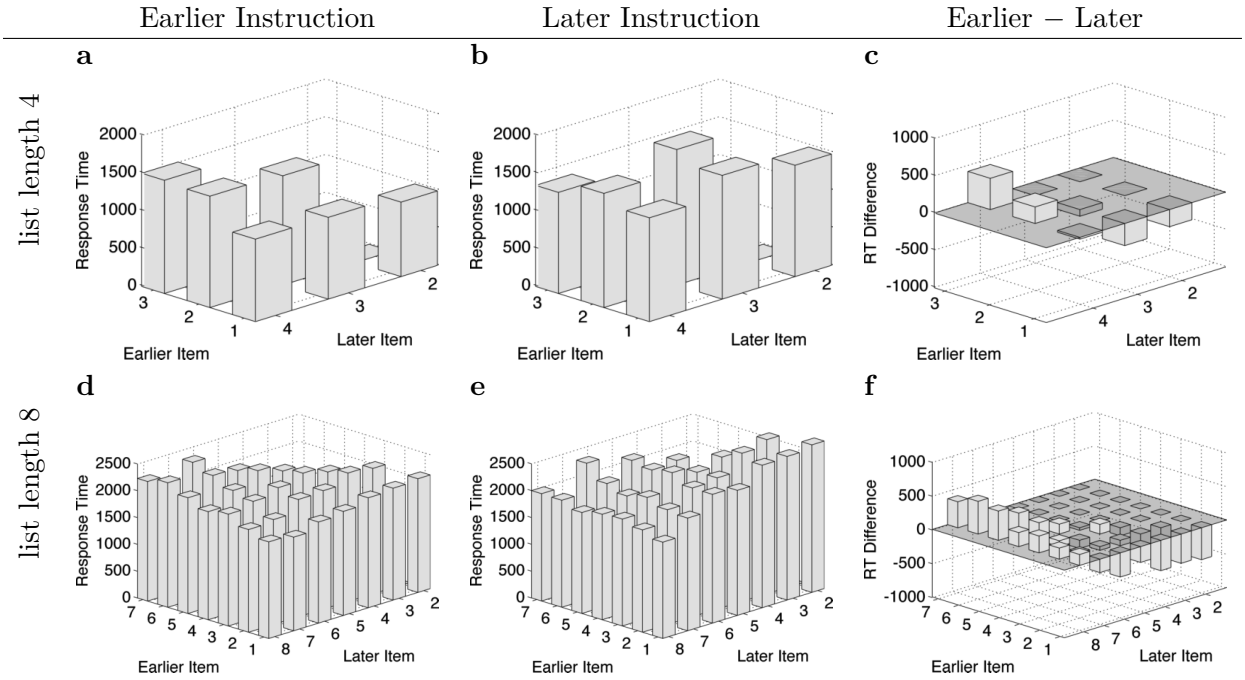

Figure 6. Response time (Experiment 2) as a function of both probe items' serial position (earlier item and later item, respectively) broken down by list length in rows, and instruction ("earlier", "later" and the difference, "earlier"–"later", corrected for mean response time) in columns.

		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	-0.032 (0.009)*
	Distance	-0.232 (0.014)*
	Later Probe Serial Position (Linear)	-29.38 (16.88)*
	Later Probe Serial Position (Quadratic)	-85.48 (7.816)*
Interactions		
	Instruction × Later Probe Serial Position (Linear)	-53.72 (4.38)*
	Trial × Later Probe Serial Position (Linear) × Instruction	-6.025 (1.45)*
	Intact/Reverse × Later Probe Serial Position (Linear) × Instruction	-109.5 (7.17)
	Distance × Later Probe Serial Position (Linear) × Instruction	-8.512 (2.085)*
	List Length × Intact/Reverse × Instruction × Later Probe Serial Position (Linear)	95.67 (5.52)*

Table 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.*

component) (Table 6). The two-way interaction is qualified by a significant four-way interaction of List Length × Instruction × Later Probe Serial Position × Intact/Reverse. We conducted additional analyses on 4 subgroups of the data: list length 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 the "later" instruction response times at Later Probe Serial Positions, and having the opposite effect on the "earlier" instruction response time at Later Probe Serial Positions. In other words, the linear Later Probe Serial Position curve associated with the "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 differences between experiments 1 and 2, including presentation rate, stimulus materials, and varied versus fixed list lengths.

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, Hulme, & Preece, 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:

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

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[1].

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

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

[1]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.

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 parameter search, we wanted to avoid finding a model that fit the "earlier" and "later" instructions individually while failing to capture the difference due to instruction. We therefore opted for a fitness measure that weighted the "earlier" data, the "later"' data and the difference pattern equally. Thus, we fitted Hacker's model by minimizing the summed BIC of the "earlier" instruction, "later" instruction and the difference between "earlier" and "later" instruction[2] (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 (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 7; cf. Figures 5 and 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 the "later" instruction. The model also cannot account for the "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 3 and 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

[2]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.

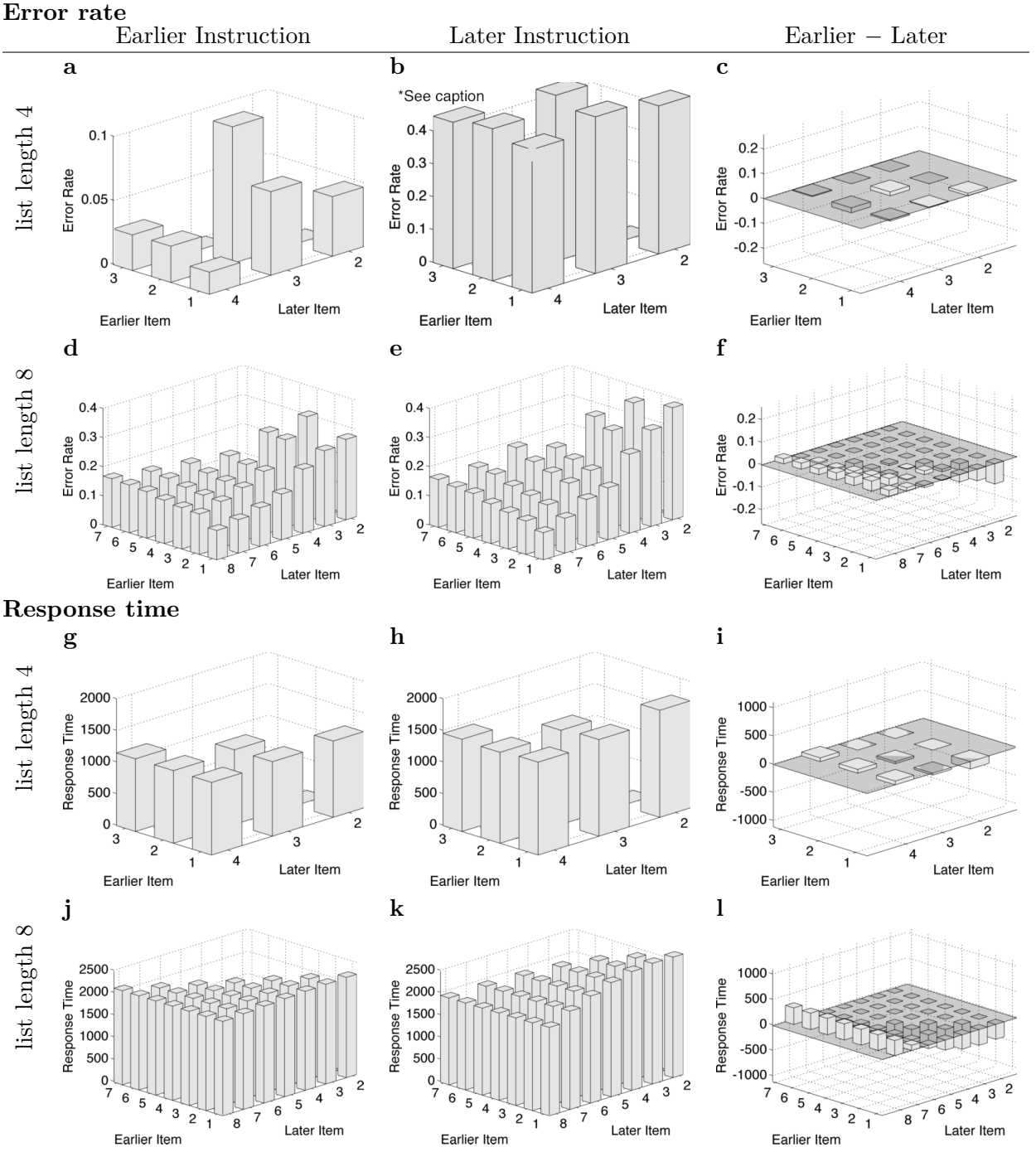

Figure 7. Hacker's model error rate (top half) and response time (bottom half), fit to experiment 2, as a function of both probe items' serial position (earlier item and later item, respectively) broken down by list length in rows, and instruction ("earlier", "later" and the difference, "earlier"–"later", 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	-8.35
list length 8	1882.04	114.96	2069.74	41.01	3.04

Table 7
Parameter summary of the Hacker forward versus backward self-terminating search model fitted for the "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 the "earlier," "later" and "earlier"–"later" instructions equally, ΔBIC in this table is computed with the "earlier" instruction data only. A negative ΔBIC indicates the forward instruction fit better.

would address this limitation.

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

To implement forward, self-terminating search, for error rate (Equation 1) we changed the first α_i to α_j:

$$P_{ij} = \alpha_j + \frac{1}{2}(1 - \alpha_i)(1 - \alpha_j) \tag{3}$$

Similarly, for response time (Equation 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 the "earlier" instruction data than the backward search model. The best-fitting model parameters from the best-fitting models are summarized in Table 7, along with ΔBIC values comparing the forward model and backward models.

The forward model fit the "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 8). Fitting the "earlier" data with the forward model and the "later" data with the backward model also improved fit of the congruity effect qualitatively (cf. Figures 5 and 6).

For more insight, note that for the forward model, the "earlier" instruction fit by decreasing α_i over serial position (Figure 9a), whereas the "later" instruction fit by increasing α_i over serial position (Figure 9b). When both "earlier" and "later" instruction fit by the backward model,

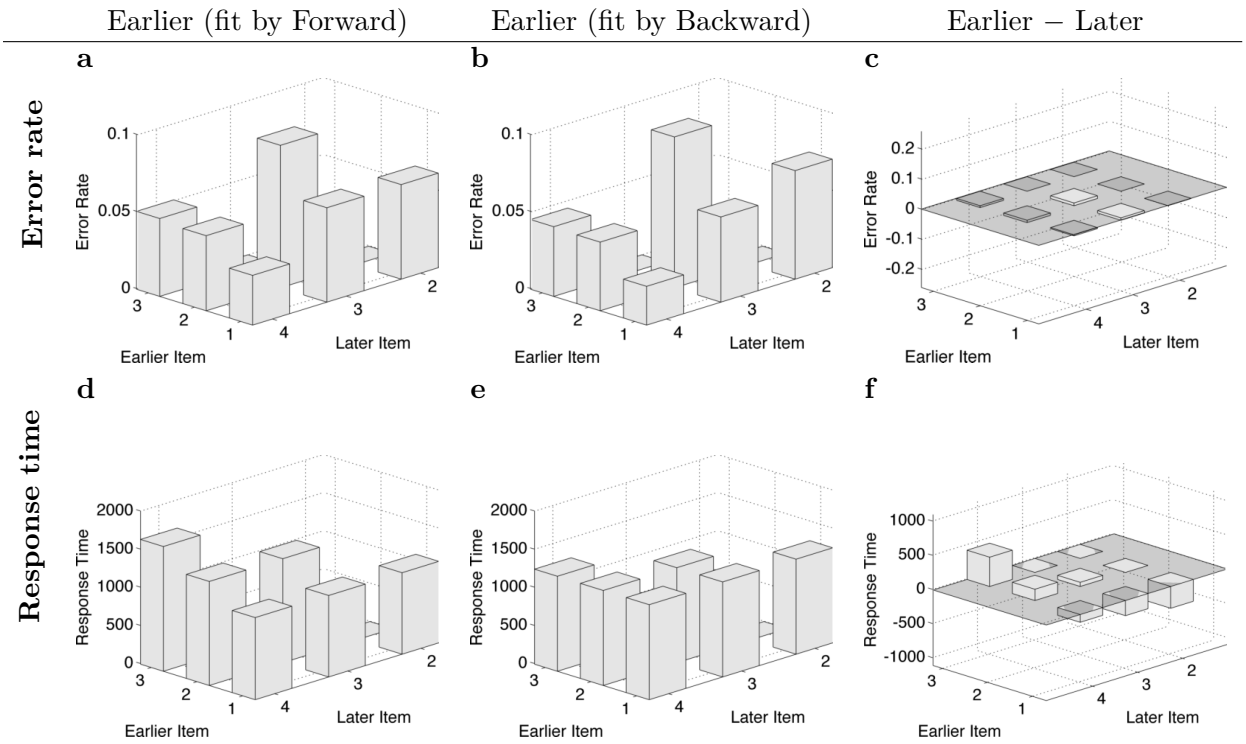

Figure 8. The best-fitting hacker's model generated plot using forward direction search for "earlier" instruction (a,d) and backward direction search for "earlier" instruction (b,e). The right-hand column (c,f) represent the hacker's model generated "earlier" – "later" difference pattern when fitting the "earlier" instruction with forward directed search and "later" instruction with the backward directional search.

the α_i values were less steeply sloped for the "earlier" than the "later" instruction. It may seem surprising that certain values of α_i were near-zero. We understand this as follows. In the "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 the "earlier" instruction and backward search for the "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

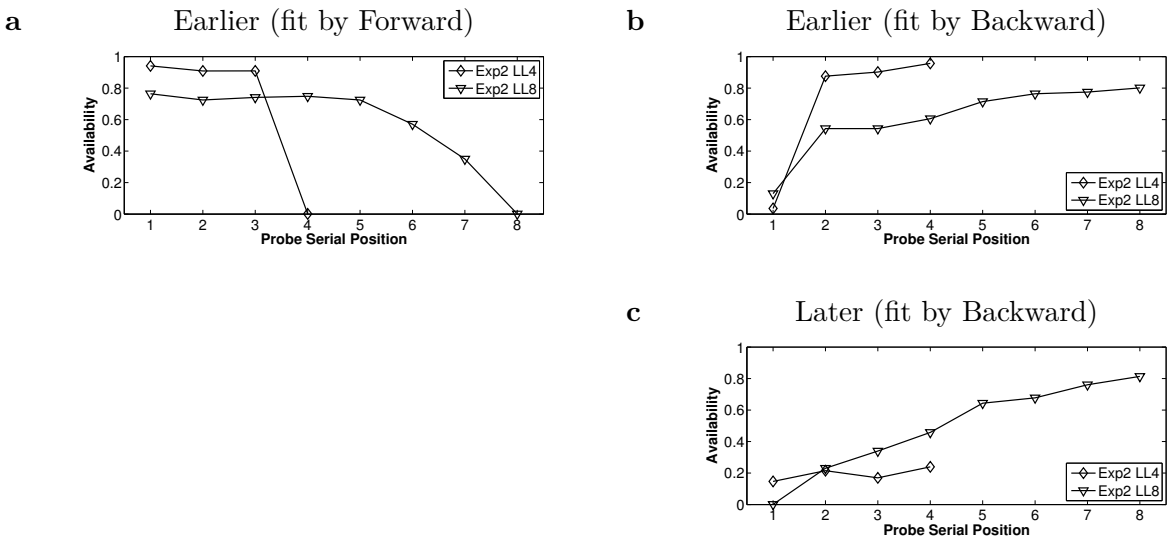

Figure 9. Availability (α_i) parameter values plotted as functions of serial position.

reasonable congruity effect. We discuss alternative model accounts in the General Discussion.

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, the "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.

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.

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 in 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 the "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 in 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 judgments.

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: the "earlier" instruction produced better judgements at earlier serial positions (primacy effect), whereas the "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 analogous dissociative effects on JOR behaviour comparing the "earlier" versus "later" instructions.

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 the "earlier" instruction and backward directional search for the "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, Preece, & Hulme, 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 the "earlier" instruction, and the last item first for the "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.

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 in 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.

References

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

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*, 12-28.

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.

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

Bates, D. M., & Sarkar, D. (2007). lme 4: 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.

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.

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

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*(3), 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.

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.

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

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

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.

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.

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.

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.

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.

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

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.

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

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

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

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.

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

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.

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

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

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.

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.

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., & 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.

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.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer*

Journal, *7*(4), 308-313.

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., 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.

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.

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

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.

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

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.

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.

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.

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 the "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 * - $p < 0.05$.*

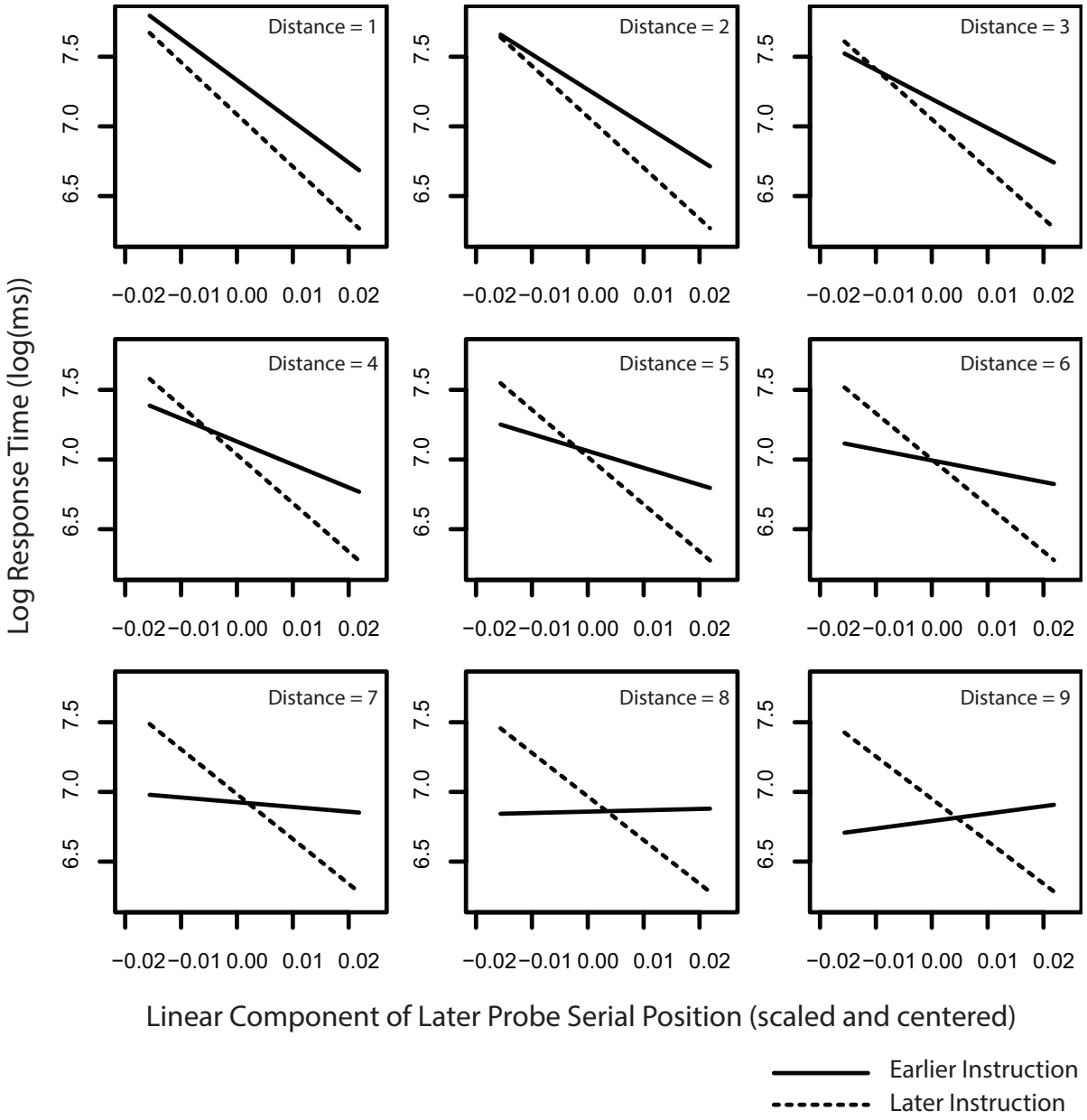

Figure S1. Best fitting LME plot of Instruction × quadratic component of Later Probe Serial Position × List Length interaction. Instruction × quadratic component of Later Probe Serial Position is plotted at all levels of List Length.

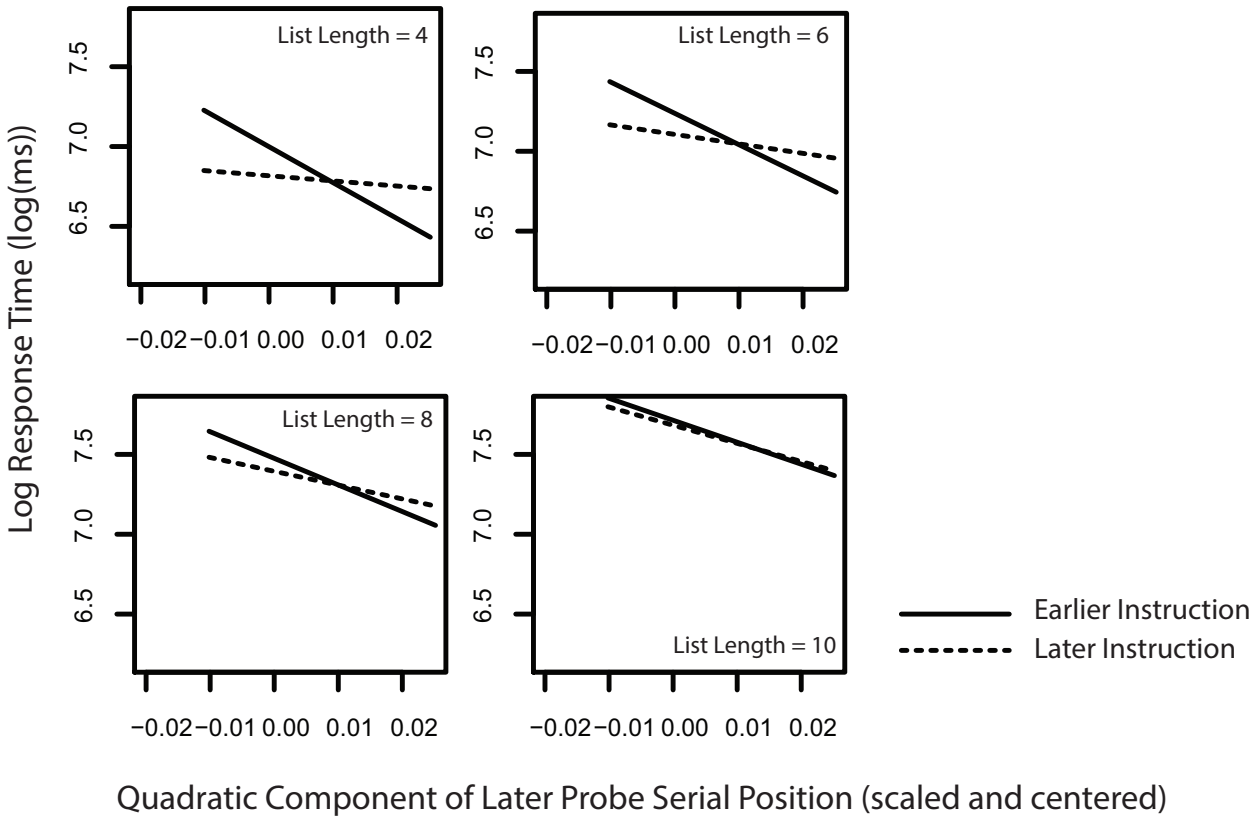

Figure S2. Best fitting LME plot of the interaction of Instruction × linear component of Later Probe Serial Position × Distance. Instruction × linear component of Later Probe Serial Position is plotted at all levels of Distance.