Judgments of alphabetical order and mechanisms of congruity effects

Yang S. Liu

Department of Psychology and Department of Psychiatry, University of Alberta, Edmonton, Alberta

Jeremy B. Caplan

Department of Psychology and Neuroscience and Mental Health Institute, University of Alberta, Edmonton, Alberta

Author Note

Portions of this work were part of the first author's dissertation (Liu, 2015). We thank the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Alberta Ingenuity Fund for funding support. Data and modelling code can be accessed at Open Science Foundation (Liu & Caplan, 2019). Correspondence concerning this article should be addressed to Y. S. Liu, Department of Psychiatry, University of Alberta, Edmonton, AB, T6G 2B7 Canada (Tel: 780-492-7872, e-mail: ly6@ualberta.ca) ORCID: Yang Liu: 0000-0003-0406-8056, Jeremy Caplan: 0000-0002-8542-9900

Abstract

The congruity effect is a highly replicated feature of comparative judgements, and has been recently found in memory judgements of relative temporal order. Specifically, asking "Which came earlier?" versus "Which came later?" facilitates response times and sometimes error rates on judgements toward the beginning or end of the list, respectively. This suggests memory judgements of relative temporal order may be part of a broader class of comparative judgements. If so, the same congruity effect should also be found with the English alphabet, despite the alphabet being a longer, semantic-memory list, with forward directional encoding. A large-sample study ($N = 340$) produced a clear congruity effect in response time and even error rate (when controlled for response time). The large number of serial positions afforded by the alphabet enabled us to test a repertoire of mathematical models instantiating four distinct mechanisms of the congruity effect, against the empirical serial-position effects. The best-performing model assumed a response bias toward a discrete set of letters conceived of as "early" versus "late." respectively, an account that had previously been ruled out for typical comparative-judgement paradigms. In contrast, models implementing congruity effect mechanisms supported for conventional comparative judgement paradigms (based on reference-point theory or positional discriminability) produced quantitatively poorer fits, with more curvilinear serial-position effects that deviated from the data. The congruity effect thus extends to long, highly directional semantic-memory lists. However, qualitatively different serial-position effects across models suggest that, despite the superficial similarity, there are probably several quite different mechanisms that produce congruity effects, which may, in turn, depend on specific task characteristics.

Keywords: Order memory, Relative order judgement, Alphabet, Congruity effect, Serial order

Judgments of alphabetical order and mechanisms of congruity effects

Significance statement

We found that people are faster judging alphabetical order of letter-pairs, depending on whether they are asked which comes earlier or which comes later. The best-fitting mathematical model explained this “congruity effect” as people having a positive bias toward letters that they think of as early in the early instruction, and late in the late instruction. This differs from causes of congruity effects in other judgements of order (animal size, random lists). Explaining congruity effects is important for understanding many behaviours that demand precise order.

Introduction

Serial-order memory is critical for a broad range of human behaviour (e.g., Lashley, 1951). One of the most direct ways 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, Ross, Earle, & Caplan, 2009; Liu, Chan, & Caplan, 2014), for reasons that will soon become clear. Memory researchers interested in the JOR procedure have made little contact with a closely related paradigm, comparative judgement, which typically examines comparisons of perceptual judgements of physical magnitudes, such as 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, White, Sturgill, & Mermelstein, 1983; Čech & 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).

Three key properties are routinely found in comparative-judgement data (see Jou et al., 2020; Petrusic, 1992; Leth-Steensen & Marley, 2000, for reviews): (a) a distance effect, whereby response time decreases as the physical or conceptual distance between the probe items increases; (b) an inverted U-shaped serial position curve for response time and error rate, 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 sometimes error rate when the wording of the question is congruent with the probe on a relevant dimension. Authors have often noted that the congruity effect is not an inevitable consequence of performing a comparative-judgement task, since with a two-item probe, the two wordings are directly related to one another; for a probe consisting of A and B, if the target is not A, it must definitely be B. This leaves no rational incentive to perform the judgement differently depending on instruction.

Both the distance effect and inverted U-shaped serial position effect have been found in judgements of relative recency (Hacker, 1980; Muter, 1979; Yntema & Trask, 1963), but only recently did Chan et al. (2009) report a congruity effect, with participants performing a JOR task on 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 judgement speed toward the end of the list. Liu et al. (2014) demonstrated that the congruity effect generalized to longer temporally presented lists (8 consonants and 4, 6, 8 and 10 nouns randomly generated and presented for study only once) and could be seen in error-rate, as well as response-time data.

Liu et al. (2014) also drew a direct parallel between the congruity effects found in episodic memory tasks and those reported in comparative judgements. However, congruity effects in short temporal lists are also explainable by directional self-terminating search for one probe only, an example of an inference-based heuristic where searching until the first of the two probe items is found in memory for the list might be sufficient to make a highly accurate 2AFC judgement, bypassing a comparison mechanism. This raises two questions, which we sought to address here: (a) How general are congruity effects? For example, the English alphabet has characteristics that are beyond the range of paradigms for which the congruity effect has been sought (with the important exception of work by Jou and colleagues, discussed shortly). (b) Do all congruity effects have the same cause? The alphabet affords a large number of serial positions, and is thus well poised to assess mathematical models of the congruity effect.

First objective: test boundary conditions of the congruity effect. Because congruity effects could be found in episodic-memory tasks as well as comparative judgements, this raises the possibility that the congruity effect is universal, spanning from temporal (episodic-memory) to semantic-memory lists to typical comparative-judgement materials. However, one characteristic might undermine the congruity effect— namely, a congruity effect might not be found for a list that was overlearned and highly directional (Zhou et al., 2006), such as the English alphabet, which has been extensively studied (e.g., Jou, 1997; Jou & Aldridge, 1999; Jou, 2010; Klahr, Chase, & Lovelace, 1983), and for which judgements of order show both distance and end effects (Lovelace & Snodgrass, 1971). If the alphabet shows a congruity effect, that would suggest a common account across paradigms (e.g., Brown, Neath, & Chater, 2007). If not, that would suggest important boundary conditions.

One suggestion of the existence of an alphabetic congruity effect in response time was reported by Jou (1997). However, this procedure was not 2AFC, but three-alternative forced choice (3AFC) and five-alternative forced choice (5AFC) judgements. For the 3AFC task, participants had to select the earliest or latest letter from among letter-triplet probes extracted from the whole alphabet. This small difference in procedure changes the nature of the task in an important way. The congruity effect for 2AFC judgements was remarkable, because if one knows which item is the earlier item, one knows (by simple elimination) that the other item must be the later item, offering no rational reason to perform the judgement differently depending on whether the earlier or later probe item is requested. Asking participants to identify the earlier item in a two-item probe is thus logically equivalent to asking for the later item. This heuristic of comparison was thought to be a special case if one of the two probe items is an end items on a list (Jou, 1997), but Chan et al. (2009) showed it was also applicable on short lists, where no characteristic distance effect was evident. In contrast, when the probe consists of three or five items, finding the earliest of the probe items is sufficient to answer which item is the earliest, but is insufficient to answer which item is the latest. A strategy that is biased toward earlier letters in the alphabet could thus be more effective, reducing the number of comparisons required overall and thus producing faster response times for the earlier instruction. Conversely, a strategy that is biased to later letters in the alphabet could likewise improve both accuracy and speed for the later instruction. Given that there is a rational incentive to use a different strategy between instructions, the 3AFC and 5AFC procedure cannot speak to whether or not participants produce a congruity effect with the 2AFC procedure, which does not demand a congruity effect.

Other alphabetical order studies used the appropriate (2AFC) procedure for this question, but did not include the full range of the alphabet and were inconclusive with respect to the presence or absence of the congruity effect for the alphabet. Jou (2003) analyzed relative-order response times involving the first 9 letters of the English alphabet and found a main effect of instruction on response time, with the Earlier instruction outperforming the Later instruction, but the congruity effect did not reach significance. In comparative judgement research, when ranges are restricted to a subset of the full range as in Jou’s procedure, the congruity effect rescales to the range of stimuli used in the experiment, or that the participant conceives as the operational range (Čech & Shoben, 1985; Čech, Shoben, & Love, 1990; Hinrichs, 1970). If this holds for alphabetical order judgements, Jou's null congruity effect finding would imply that even if the full range of the alphabet were used, as we do here, the congruity effect should be expected to be absent. For example, the participants may either perceive the list as a list of 9 items consisting of the first 9 letters from the alphabet, or a list of 26 items consisting of the full range of the alphabet. Alternatively, if an overlearned set such as the alphabet is less amenable to range-rescaling, it could be that participants operate on a hypothetical continuum of the full 26 letters of the alphabet. If participants consider the set as a list of 26 items, the "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, put differently, the congruity effect may be easy to observe across the whole list, but may become quite subtle at the edges. We revisit this finding in the Discussion and Model sections. Jou and Aldridge (1999) tested more of the alphabet (excluding the first and last three letters), but they did not analyze the congruity effect (serial position was not included in their ANOVA). Based on the published results, we cannot know whether a congruity effect would have been significant or not.

These failures to observe a significant congruity effect might indicate that there is no congruity effect in alphabetical order judgements. Liu et al. (2014) proposed that the congruity effect may reflect participants in one group (Earlier instruction) processing the list in the forward direction, and in the other group (Later instruction) processing the list in the backward direction. Given that the alphabet is strongly forward-directional, participants may not be able to flexibly reverse their direction of processing. In this case, no 2AFC congruity effect would be expected for the English alphabet. If, indeed, a null congruity effect were found for the alphabet, forward-directionality may be an important boundary condition on congruity effects.

On the other hand, there are good reasons to predict that with sufficient power, one should observe a congruity effect in JORs of the English alphabet. Given that the alphabet could be considered to reside in semantic memory, semantic-memory congruity effects have 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-judgements of the events in a story script (Wyer, Shoben, Fuhrman, & Bodenhausen, 1985) and relative-order judgements of autobiographical episodes (Fuhrman & Wyer, 1988). For a longer semantic list, such as the English alphabet, we may observe a similar response-time congruity effect.

The experiment we report here differs from the methods of Jou (2003) and Jou and Aldridge (1999) in two ways. First, it tests the full range of the alphabet; if the congruity effect is robust for probes at the very start and very end of the alphabet, this may increase sensitivity to a congruity effect. Second, we collect more data (340 participants and 650 trials per participant here, compared to 62 participants and 84 trials in Jou, 2003 and 38 participants and 380 trials in Jou & Aldridge, 1999). A precise comparison of sensitivity is difficult due to numerous unknown factors. For example, we manipulated instructions between subjects, whereas both prior studies manipulated instructions within-subjects. Our analyses may lose sensitivity due to subject variability. On the other hand, when conducted within subjects, participants might perform the two instructions more similarly to one another, reducing any instruction effects. In addition, we use linear mixed effects (LME; Baayen, Davidson, & Bates, 2008; Bates, 2005) to analyze the data; by including additional variables in the model, we may have greater sensitivity to an underlying congruity effect than the two prior articles that used conventional ANOVA.

Second objective: test model mechanisms of the congruity effect. The field of comparative judgement research has an extensive record of testing model-mechanisms of the congruity effect. Our second objective was thus to test whether the favoured models for comparative-judgement data generalize to alphabetical-order judgements. We also consider models that are not thought to characterize congruity effects in typical comparative-judgement tasks.

First, we note that much work in comparative judgement research has gone into determining the nature of the basic comparison process, with a focus on whether judgements are based on discrete or semantic codes versus continuum-based or magnitude judgements. However, as Petrusic, Shaki, and Leth-Steensen (2008) noted (and see Čech, 1995; Jou, Escamilla, Torres, Ortiz, & Salazar, 2018; Jou et al., 2020; Moyer & Dumais, 1978; Shaki & Algom, 2002), this can be orthogonal to determining the mechanism of the congruity effect itself. Put differently, the congruity effect may emerge during the comparison process itself, or the congruity effect may be an additive step, either preceding or following the comparison. Next, we very briefly note several classes of accounts of the congruity effect derived from the comparative judgement literature.

Serial, self-terminating search. In the domain of episodic memory research on short-list judgements of relative order, Sternberg (1969, Experiment 8) saw his findings as evidence that participants performed serial, self-terminating search, later corroborated by Hacker (1980). Building on this, Chan et al. (2009) proposed that their short episodic memory list congruity effect was due to a reversal of search direction, with participants searching forward in the "earlier" instruction and backward in the "later" instruction, similar to an idea proposed by Moyer and colleagues (Moyer & Bayer, 1976; Moyer & Dumais, 1978). Liu et al. (2014) showed that a numerical implementation of such a model, adapted from Hacker's model, fit new data from short lists well. Meanwhile, subsequent to Moyer and Bayer (1976), comparative-judgement researchers have either not considered serial, self-terminating search or consider this account incompatible with distance effects or response times that do not increase linearly with serial position (Banks, 1977a; Lovelace & Snodgrass, 1971). Unlike short, episodic memory lists, a reversal of search direction has not been favoured for standard comparative judgement paradigms.

Semantic-Coding Theory. Banks, Clark, and Lucy (1975) proposed that stimuli are judged by retrieval of a categorical label (elaborated by Banks, 1977a). Applying this to the alphabet, "early" letters would have the code E, and "late" letters, L. The assumption is that participants retrieve the category label of the two items. If one is E and the other L, the judgement is made directly. If both are E, then a finer distinction must be made to determine E+ (more "early") versus E (less "early"). If the instruction asked the participant to select the earlier item, the response can already be made; if the later item is required, the participant must either reverse or recode E+, E to L, L+, resulting in an additional cost to response time for incongruent trials. Semantic-coding theory has been challenged. First, non-verbal primates exhibit congruity effects (Cantlon & Brannon, 2005; Cantlon, Platt, & Brannon, 2009), and perceptual judgements also exhibit congruity effects (e.g., Petrusic, 1992; Petrusic & Baranski, 1989). Thus, linguistic codes, or even language, cannot be essential to produce a congruity effect. The converse, however, is possible: congruity effects might derive from linguistic codes when those are available. A second line of argument against a semantic-coding mechanism regards findings of interactions of the congruity effect with probe difficulty, which are incompatible with semantic-coding (e.g., Chen, Lu, & Holyoak, 2014). A third line of argument is that the semantic congruity effect is presumed to happen at the semantic coding translation stage, thus cannot explain empirical findings that the use of nonsense consonant-vowel-consonant syllables to replace conventional instruction can enhance the congruity effect, since this occurs prior to that translation (Petrusic et al., 2008).

Expectancy theory and Semantic Interference theory. A close cousin of semantic-coding theory is expectancy theory (Marshuetz, 2005; Marschark & Paivio, 1979, 1981). The idea here is that while anticipating the probe, participants prime or somehow pre-activate stimuli that are congruent with the instruction. This has been severely challenged by findings of congruity effects even when the instruction *follows* the probe (Holyoak & Mah, 1981; Marschark & Paivio, 1979, 1981; Shoben, Čech, Schwanenflugel, & Sailor, 1989), expanded on by Jou et al. (2018). Again, this may rule out expectancy as the only mechanism of congruity effects, but does not rule it out as one of several possible mechanisms, when available.

A related account, semantic interference theory (Banks & Root, 1979), compares the congruity effect to the Stroop effect (but see Shaki & Algom, 2002). This raises the possibility that a Stroop-like mechanism could potentially produce a congruity effect even without linguistic processes (non-semantic). In other words, when choosing the earlier target, but when the target is considered a “Late” item, or choosing the later target from amongst two “Early” items, there is response conflict that needs to be overcome, lengthening response times. An important point is that this type of mechanism, where response times are lengthened when probe-item class conflicts with the target being sought (Late items in Earlier judgements or Early items in Later judgements), could plausibly occur at the response phase, which differentiates semantic interference theory from expectancy theory.

Reference-point or anchor theory. The majority of the mechanisms of the congruity effect that are currently favoured for comparative judgements are, upon close inspection, forms of reference-point or anchor theory, including evidence-accrual theory (e.g., Chen et al., 2014; Holyoak, 1978; Jamieson & Petrusic, 1975; Leth-Steensen & Marley, 2000; Marks, 1972). At the core of these accounts is the idea that prior to comparison, magnitudes are computed or derived from distances to a reference point (or similarities to a reference point; Leth-Steensen & Marley, 2000) which is usually the start or end of the stimulus set. Given a nonlinearity (in our own model, we use a log transform, in line with SIMPLE; Brown et al., 2007), this results in greater discriminability between probe pairs that are closer to the reference point than those that are farther away. A congruity effect arises when the instruction induces participants to use the start versus the end of the list, respectively, as the reference.

Positional distinctiveness of the target. With a series of clever experiments, Jou et al. (2018, 2020) suggested that the congruity effect in comparative judgements may originate due to items being facilitated (perhaps akin to priming or pre-activating) in proportion to their positional distinctiveness. In their account, congruity effects arise simply because in one condition, the target item will tend to be closer to one end of the set. This increases discriminability of the outer target, and discriminability drives the decision. The instruction simply changes which of the pair of items is the response. This is a compelling idea, because it does not require language, expectation or adjustment during study, nor a difference in processing the probe. It is also compatible with findings of congruity effects in paradigms for which the instruction follows the probe.

Goals of the experiment and models

To test the generality of the congruity effect and compare model mechanisms, we designed an experiment examining relative-order judgements of letters from the English alphabet. Both response-time and error-rate were measured. We tested the full range of the English alphabet, with all possible probe combinations (with equal probability), and manipulated instruction between subjects. We tested a large sample ($N = 340$, 650 trials per participant) because we were curious about the possible presence of error-rate effects, that should be quite subtle given that accuracy was expected to be near ceiling. In Liu et al.'s (2014) Experiment 2, 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 in error rate. We therefore aimed for nearly twice the number of participants here.

Experiment: Methods

Participants

The research was approved by the University of Alberta's Human Research Ethics Board (Pro00014801). 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 six. 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 condition cycled across groups in order of arrival. Twelve participants were excluded because their accuracy suggest low engagement of the task (below 80%), or they self-reported having not followed the instructions. Final analyses thus included 173/175 and 155/165 participants in the “earlier” and “later” groups, respectively.

Materials and Procedure

The experiment was created and run using the Python Experiment-Programming Library (Geller, Schleifer, Sederberg, Jacobs, & Kahana, 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 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 centre of the screen, followed by the probe consisting of two letters from the English alphabet, after which the participant made their response by pressing the ‘.’ key (for the left-hand probe item) or the ‘/’ key (for the right-hand probe item) on a QWERTY keyboard. 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).

LME analysis was applied to our data to determine how instruction affected error rates and response time (Baayen et al., 2008; Bates, 2005). 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"), linear component of Serial Position (linear component of the serial position of the probe item that appeared earlier from the presented list), quadratic component of Serial Position, Distance (absolute value of the difference between the serial positions of the two probe letters) and Intact/Reverse (whether the probe order was consistent/inconsistent with presentation order, respectively) were included as fixed factors. Because serial position effects were expected to have a quadratic-like form with performance better both at early and late than middle serial positions, we included the quadratic component in the data model. The quadratic component is non-orthogonal to the linear component, thus not allowed to interact with the linear component in subsequent analyses. Subject was included as the only 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 attenuate skewness. Error rate data were fitted with logistic regression, suitable for binary variables ("correct"/"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), as done by Liu et al. (2014), using LMERConvenienceFunctions (Tremblay, 2013).

Experiment: 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 the "earlier" and "later" instructions, respectively), which were significantly different, $\chi(1)^2$=35.32, $p < 0.001$. Thus, by this very coarse measure, the "later" instruction appears to have been challenging. Corroborating this, without any exclusions, the error-rate difference, 0.020, was significant ($t(265) = 2.58$, $p = 0.010$, independent-samples). Correct-response time did not differ significantly, $t(338) = 0.31$, $p = 0.756$.

Next, we looked to the LME results to find out if there was a significant congruity effect, characterized by a crossover interaction between Instruction and the linear or quadratic component of 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 and Figure 4). The instruction by serial position interaction and the "earlier"–"later" mean difference are plotted for response time (Figure 1) and error rates (Figure 2).

Response Time. The response-time congruity effect can be clearly seen (Figure 1)— 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 (linear component) interaction in the best-fitting LME model (Estimate = -0.136, $p < 0.05$; Table 1). The difference in mean response time between "earlier" and "later" instruction shows an approximately linear increase from the earlier-probe serial position 1 to 25. When Instruction = "later" the effect of the linear component of Serial position is –0.136 units lower than when Instruction = "earlier." Note that the congruity effect is of the same order of magnitude as the quadratic component of the serial position effect (Estimate = -0.168, $p < 0.05$), one of the most robust findings in serial-order memory research.

The Instruction × Serial position (linear) interaction was also part of higher-order interactions, including interacting with Intact/Reverse (Estimate = -0.018, $p < 0.05$), Distance (Estimate = -0.032, $p < 0.05$) and Trial (Estimate = -0.008, $p < 0.05$). However, in all these higher-order interactions, the form of the two-way interaction was the same, and therefore do not alter the conclusion that the congruity effect was indeed present.

Error rate. Inspection of the plot of Instruction × Serial positions (Figure 2) suggests no congruity effect. However, when we compared the error-rate data (Figure 2) to the response-time data (Figure 1), this suggested to us that the "earlier"−"later" instruction difference at higher serial positions shows an opposite pattern for response times and error rates. 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 the "later" instruction, with the side-effect of producing more errors late in the alphabet in that group. In other words, many of the faster responses may have been errors, and more so for the "later" group than the "earlier" group. Therefore, if speed were controlled for (on both correct and error trials), we could then test whether error rate showed an underlying congruity effect. Including response time in the LME model for error rate (Table 2), the best-fitting model failed to converge using the LME4 package. We thus repeated model selection using logistic regression with the same model, without adding Subject as a random effect and using the glm() and step() functions in R. Other than the significance of the Instruction × Serial position (linear) interaction (Estimate = –0.081, $p < 0.05$, the resulting best-fitting model was consistent with the original LME best-fitting model (without response time as a predictor) as to which effects were significant, as well as the directions of the effects. Both logistic regression (Table 2) and LME (Table S1) best-fitting models showed a significant Instruction × Serial position (quadratic component) interaction (Estimate = 0.215, and 0.224, respectively, $p < 0.05$), confirming a congruity effect with the error-rate measure.

Our exclusion criteria using 200 ms as a lower cut-off and above 4 standard deviation as an upper cutoff for response times are standard quality-control measures to ensure we are measuring valid behavioural data and not artefact or behaviour of participants who are not engaging in the task we wish to study. To check the robustness of our results to these data-analysis choices, we re-ran the analysis with all data included and found the magnitude and direction of congruity effect in both correct response time (Estimate $= -0.126$, $p < 0.05$ versus Estimate $= -0.136$, $p < 0.05$) and error rate (Estimate $= 0.413$, $p < 0.05$ versus Estimate $= 0.224$, $p < 0.05$) are consistent (see corresponding effects in Tables 1 and 2).

Discussion of Experiment

We found a congruity effect in 2AFC judgements of an over-learned, highly directional list, the English alphabet. The congruity effect was observed for response times, and also in error rates, although only after controlling for speed-accuracy trade-offs. Both congruity effects have not been previously reported in alphabetical order judgements. We deliberately collected a large sample, more than five times the sample size of Jou and Aldridge's (1999) ($N = 38$) and Jou's (2003) ($N = 62$) studies, which did not report a significant congruity effect, suggesting prior studies had insufficient power (Jou, 2003) or did not test for it (Jou & Aldridge, 1999). This builds on comparable findings with subspan (Chan et al., 2009) and supraspan (Liu et al., 2014) temporal-order JOR tasks. The congruity effect may therefore be a benchmark phenomenon for order-memory judgements, shared across a broad range of list lengths, and for semantic-memory lists as well as temporally presented lists. Because our experiment had no "study" phase, the congruity effect cannot be explained by differences in encoding processes, as is possible for episodic-list judgements of relative order (Chan et al., 2009; Liu et al., 2014). It follows that at least some portion of those prior congruity effects may occur at time of test, as Hintzman (2016) wondered. The congruity effect for alphabetical order judgements is thus in good company, along with comparative judgement tasks, which either test pre-experimental knowledge (e.g., animal sizes) or newly trained memory sets.

In comparative judgement research, instruction is nearly always manipulated within subject, usually post-study and sometimes even post-probe (Jou et al., 2018). When blocked, rather than mixed from trial to trial, congruity effects tend to reduce in magnitude (e.g., Čech, 1995; Marschark & Paivio, 1979; Shaki, Leth-Steensen, & Petrusic, 2006). Because instruction in our experiment was not just blocked but also manipulated between subjects, it is plausible that we are underestimating the congruity effect we would obtain if instruction varied within subjects.

In addition, Jou (2003) limited the probes to the first 9 letters from the alphabet and reported a main effect of instruction, with no interaction between instruction and serial-position. As elaborated in the Introduction, the main effect over the first nine letters could be a sign that participants do not rescale the comparison continuum to the range of experienced stimuli, in the case of the alphabet— that is, they always conceive of the entire alphabet as the functional stimulus set, even when tested only on a subset. If this were the case, we should be able to replicate Jou's main effect but lack of crossover interaction within the first nine letters. Thus, we filtered our data set to include only probes from the first 9 letters of the alphabet (Figure 5a). We also examined the data filtering only the middle 8 letters of the alphabet (Figure 5b), and only the last 9 letters of the alphabet (Figure 5c). We obtained the best-fitting LME model and found main effects of Instruction and Instruction × Serial Position (linear) for the first 9 (Estimate = –0.064, $p < 0.05$; Table S2) and last 9 (Estimate = –0.120, $p < 0.05$; Table S3) group, but not for the middle group (Table S4). The best-fitting LME models for probes within first 9 letters of the English Alphabet does show a significant Instruction by Serial position (linear component) Interaction ($\Delta BIC \leq 2$, $p > 0.05$). However, visual inspection of Figure 5a shows that this was not a crossover interaction. Thus, increased sensitivity notwithstanding, this is in line with Jou's (2003) dominant main effect of Instruction (faster response time for the earlier instruction), if we assume that the alphabet list is always conceived as a set of 26 items, regardless of the JOR testing range. Further, we found the opposite pattern when filtering the data to the last 9 letters of the alphabet (Figure 5b), where the Later instruction had a faster overall response time than the Earlier instruction (Estimate = –0.073, $p < 0.05$), and we found a small non-crossover Instruction × Serial position (linear) interaction (Estimate = –0.120, $p < 0.05$). No main effect of Instruction nor Instruction × Serial position was found for the middle letters ($\Delta BIC <= 2$, $p > 0.05$) (Figure 5c).

An error-rate congruity effect, to our knowledge, has not previously been reported for the English alphabet. We also might have overlooked this effect had we not (after data-collection) considered that a subtle sort of speed–accuracy might be at play. 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 Liu et al.'s (2014) supraspan JOR results. It may seem odd to presume that response time in some way causally precedes accuracy, particularly if both errors and response latencies are produced by a single comparison process, which some theories assume. However, if the congruity effect occurs after the comparison process has been made, then it could be the case that the congruity effect induces participants to rush, at the expense of more errors. Considering that the beginning of the alphabet is better mastered than the end, rushing to respond prematurely at the end of the alphabet should be more detrimental than rushing at the beginning of the alphabet. If the congruity effect emerges because the earlier instruction urges participants to respond faster when the probes are early and the later instruction, when the probes are late, error rates for the early letters and late letters will be affected differently. For example, if both probe letters are "late" letters, and both response options are be primed to the same degree by the Later instruction, the participant would be induced to respond quickly, at the expense of producing more errors. The Earlier instruction would not have primed those Late letters, avoiding the inducement to rush to respond. If the Later-participant overcomes the urge to respond too quickly, that error-rate penalty might be neutralized, which is why controlling for response time may have revealed a congruity effect in the error-rate measure. We revisit this in the General Discussion, after showing how the model-selection leads us to favour such a post-comparison mechanism. In contrast to the robust finding of the response-time congruity effects, error-rate congruity effects are more unusual in comparative judgements. One possible explanation is that the error-rate congruity effect is very subtle and a large N and large number of trials per condition per subject are required to detect this effect. In addition, with tasks involving materials that are overlearned, researchers typically focus, understandably, on response-time as the principal behavioural measure (e.g., Birnbaum & Jou, 1990). One notable exception is an error-rate congruity effect found with 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 (Petrusic, 1992).

Models of the Congruity Effect

So far our empirical findings suggest that the congruity effect is quite general across comparative-judgement and memory paradigms. However, multiple model mechanisms have been proposed to underlie congruity effects. In this section, we were interested in whether the congruity effect we reported for the alphabet might have a cause that is common to congruity effects found in other tasks. The large number of serial positions afforded by the alphabet gives us the opportunity to look to the forms of serial-position effects to evaluate models.

Here we simulate very simple quantitative models of the congruity effect and evaluate them against the serial-position effects found in our data. Because the models are simple enough to have three or fewer free parameters, we were able to perform direct searches to optimize their fit to the data, quantified by the Bayesian Information Crtierion (BIC). In addition to quantitative fit, we also examine qualitative criteria of the best-fitting model. A model will be considered to be a good characterization of alphabetical judgement behaviour if it shows (a) an approximately linearly increasing function across serial position (reflecting an interaction) when all probe combinations are analyzed, (b) main effects of instruction when the only the first 9 or only the last 9 letters are considered (and no crossover interaction with serial position), and (c) nearly no effect of instruction when only the middle letters are considered.

Models of serial-order memory have not yet been designed to produce congruity effects. The congruity effect may be diagnostic of memory models or suggest how existing models might be further developed (Liu et al., 2014), and we consider one such model, SIMPLE (Brown et al., 2007). Theories designed to explain comparative judgements may also shed light on our understanding of congruity effects across cognitive domains. We consider several of these.

First, we do not consider serial, self-terminating search (with a reversal of search direction between instructions), as it is hard to reconcile with the distance effect and the serial-position effects not increasing linearly across the alphabet. We consider the following models (a) Categorical Interference model, a mechanism that assumes two categories of letters, and an interference mechanism akin to Banks and Root (1979), but not necessarily verbal. The two-class category notion is similar to Semantic-Coding theory (Banks, 1977a; Banks et al., 1975; Banks, Fujii, & Kayra-Stuart, 1976), but again, need not be semantic. It is similar to Expectancy theory, but could just as plausibly take place following the probe. (b) Anchor-point or reference-point model (Holyoak & Mah, 1981). (c) Positional distinctiveness theory (Jou et al., 2018, 2020). (d) A version of SIMPLE with an added linear gradient of discriminability, adapted from a model by Liu (2015).

Note that we are not modelling the full alphabetical-order judgement task. We are only concerned with the source of the congruity effect, so our models are greatly simplified to target the congruity effect alone. Thus, any of these models of the congruity effect might plausibly co-exist with any full model of the alphabetical judgement task.

The MATLAB/Octave code for all models, as well as an all-purpose plotting script, are included in the Open Science Foundation project (URL:

`https://osf.io/fsc8r/?view\_only=983f87c245f548f9b9c8f159c0b25fdd`) (Liu & Caplan, 2019).

Direct search

For each model, all combinations of the free parameters considered were tested by simulating or solving the model for all combinations of the earlier and later probe serial positions. BIC was computed from the root-mean-squared deviation of the best-fitting model values (inverted so that higher values transform into lower response times) from the observed means (Earlier–Later response times). Table 3 displays, for each model considered, the free parameters along with the range searched and step size, and reports the best-fitting parameters and BIC for the best-fitting model. Following convention, two models were considered meaningfully different if $\Delta BIC > 2$.

1) Categorical Interference Model

This mechanism embodies one core idea that is behind Semantic Coding theory and Expectancy theory, but without committing to verbal codes, nor to whether there is any role for expectancy. Namely, we keep only the idea that there are two classes of list items (letters), and that, completely independent of the comparison process, itself, the instruction leads letters belonging to the instruction-congruent class ("Late" letters for the "Later" instruction and "Early" letters for the "Earlier" instruction) to be responded more quickly as targets. Equally plausible, incongruent letters might be inhibited as responses and thus delay response times as the participant overcomes this inhibition. Seeing no way to distinguish these two possibilities with our data, our model is ambiguous with respect to the mechanism being facilitative, inhibitory, or both. To implement the idea of a two-class category, we assumed that participants conceptualize (either verbally or non-verbally) letters at the start of the alphabet as "Early" and letters at the end of the alphabet as "Late." Next, we simply assumed that the "Earlier" instruction would facilitate probes for which the target was from the Early set, and the "Later" instruction would facilitate probes for which the target was from the Later set. We assumed that the amount of facilitation would be equal and not further dependent on serial position in any way. Specifically, let θ be the cut-point between Early and Late letters and i, j index letter rank-position of the two probe letters (ranging from 1 to 26). Let $E(i)$ and $L(i)$ be functions for Earlier and Later instructions, respectively, that represent the relative amount of facilitation (negative values, shortening response times) due to the compatibility of the instruction with the letter:

$$E(i) = \begin{cases} -1 & i \leq \theta \\ 0 & i > \theta \end{cases} \qquad L(i) = \begin{cases} 0 & i \leq \theta \\ -1 & i > \theta \end{cases} . \tag{1}$$

The magnitude of the congruity effect, $C(i, j)$, for a given probe (i, j) is computed:

$$C(i, j) = E(i) - L(j) \; \forall i < j, \tag{2}$$

This model had only one free parameter, an overall scaling factor by which the model data were multiplied to enable it to align with the data (Table 3). As expected, this model, with $\theta = 13$ (letter M), naturally produces a dominant main effect favouring the Earlier instruction when only the first 9 letters are considered (Figure 6b), and a main effect favouring the Later instruction when only the last 9 letters are considered (Panel d). It produces a monotonically increasing congruity effect across the middle positions (Panel c) as well as across the entire alphabet (Panel a). However, the most obvious way in which the model misses the data is in the sharp discontinuity at the categorical cut point (position 13).

We considered that different participants might use a different cut point between Early and Late letters. To test the effect this might have, we averaged the serial-position effects over every cut point across the entire range to simulate the simple assumption that all cut points over the range A–Y were equally probable. Although unlikely to be strictly true, this kept the model parsimonious. Like the cut-point 13 model, this model still had only one free parameter, a scaling factor (Table 3). Figure 7 shows that the model comes close to the desired main effect for the first-9 and last-9 probes (but slightly increasing with serial position, following the data), and small interaction for middle-letter probes (Panels b–d), and is no longer jagged across all positions (Panel a). BIC was far lower than the model with the fixed cut-point at letter M.

Given the improvement when we assumed a uniform distribution of cut-points across participants, we wondered if the model would fit even better if it were assumed that cut points vary only within a particular range rather than A–Y. We added two free parameters to the search. This model (not plotted) fit best when the range of cut points was D–W (Table 3) but evidently, the better numerical fit was offset by the addition of two free parameters; the BIC was substantially lower than the model assuming a uniform distribution of cut points.

2) Reference Point Model

Next, we implemented a simple version of the Reference Point model, as described by (Holyoak & Mah, 1982). We assume that the positions of items are evaluated relative to a reference point, either $r_E =$ position zero (one letter-distance earlier than the letter A), used in the Earlier instruction, or $r_L =$ position 27 (one letter-distance later than the letter Z), used in the Later instruction. Reference-point and anchor-point models assume that values along a continuum are more discriminable when those values are closer to a reference point. Borrowing from distinctiveness-based models (e.g., Brown et al., 2007), we implement this by assuming that positional distances are log-transformed:

$$e(i) = \log(i - r_E), \tag{3}$$

$$l(i) = log(r_L - i). \tag{4}$$

$E(i, j)$ and $L(i, j)$ are simply:

$$E(i,j) = e(i) - e(j), \tag{5}$$

$$L(i,j) = l(i) - l(j). \tag{6}$$

The congruity effect is computed as the difference between the positional distance of the target item from the reference point, for any pair of serial positions:

$$C(i,j) = E(i,j) - L(i,j). \tag{7}$$

Searching two free parameters, the base of the logarithm and an overall scale factor (Table 3), this model (Figure 8) does a good job of producing relatively flat serial-position effects for subsets of the alphabet (panels c, d). However, it underestimates the size of the congruity effect in this dataset in the first and last nine subsets. More seriously, this model shows the congruity effect accelerated toward the reference points; in this figure, this effect is evident toward the start of the alphabet (panels a and b), whereas the effect is evident toward the end of the alphabet when plotted as a function of the later-probe serial position (not shown). This, combined with the overall serial-position effects being more curvilinear than the data (panel a) along with the poor BIC value, suggests that the Reference Point model is not a good account of the current data.

3) Positional Distinctiveness model

As described in the introduction, Jou et al. (2018, 2020) proposed that list items are advantaged due to their positional distinctiveness, and all that the instruction does is specify which of the probe items is to be the target. Targets that are closer to the ends of the set (earlier items, for the "earlier" instruction, and later items for the "later" instruction) will receive a processing advantage. Jou et al. (2018) calculated positional distinctiveness for each item to each other item (their Table I.1) and found it to be a good fit to their free-choice data. We take this one step further, and compute the congruity effect itself, subtracting positional distinctiveness values, $D(i)$, for all probe combinations, depending on whether the earlier or later item is the target. We first take the log of inter-item distances, to be consistent with Murdock (1960), whose model inspired this one:

$$D(i) = \sum_{k=1,\ k \neq i}^{L} log(i - k). \tag{8}$$

The congruity effect is:

$$C(i, j) = D(i) - D(j),\ \forall i < j. \tag{9}$$

In other words, this model estimates the congruity effect by, for each possible probe, subtracting the distinctiveness of what would be the target in the Earlier instruction (the earlier probe-item) from what would be the target in the Later instruction (the later probe-item). Searching two free parameters, the base of the logarithm and an overall scale factor (Table 3), this model (Figure 9) produced a pronounced curvilinear serial-position effect (panel a), clearly deviating from the data, which are closer to linear. The model did produce an approximate null effect when only middle-position probes were considered (panel c), resembling the data, but underpredicted the magnitude of the congruity effect when restricted to the first (panel b) or last (panel d) letters of the alphabet. Along with a relatively poor value of BIC, this model seems not to be a good fit to the data.

4) Gradient of Discriminability model

Liu (2015) proposed that SIMPLE, a scale-invariant memory model based on relative temporal discriminability (Brown et al., 2007), could be modified to produce a congruity effect by assuming that the representation of time was “stretched,” with the stretching more pronounced toward the start of the list for the Earlier instruction, and toward the end of the list for the Later instruction. If one considers that reference points can enhance discriminability at one end versus the other of a dimension (Chen et al., 2014; Holyoak &

Mah, 1982), this mechanism could be viewed as an instantiation of Reference Point theory. The model provided a good account of episodic memory judgements of relative or when lists were long enough (above list length 4), but not when they were short (Liu, 2015). The Start-End Model (SEM; Henson, 1998) also assumes memory is driven by positional distinctiveness, but computes distinctiveness differently than SIMPLE. The SEM assumes each item is associated to a strength relative to the start and to the end of the list, respectively. It produces many of the features that SIMPLE also produces, such as bow-shaped serial-position effects. SEM would be an interesting alternative to SIMPLE, particularly given that it is one of the handful of models that have been fit to a broad range of empirical benchmarks. However, due to its large number of free parameters, parameter-optimization is more challenging. For this reason, we instead considered SIMPLE, but we expect that a similar adaptation of SEM would produce similar results, given that they share the basic positional-distinctiveness mechanism. Here we implement the same mechanism, but because the alphabet is learned over many sessions, "recency" is not relevant, so we treat the list as Brown et al. (2007) treat identification tasks, and use positions 1–26, not log-transformed. Unlike Liu (2015), who implemented SIMPLE fully, here we isolate the portion of the model that could explain the congruity effect (Earlier–Later condition). Specifically, let parameter $g = 0.01$ set the magnitude of the linear gradient across position, and let $c = .1$. We define $p_e(i)$ and $p_l(i)$ as the position codes for the Earlier and Later instructions, respectively:

$$\text{and } p_l(i) = i(1+g), \tag{10}$$

$$p_e(i) = i + (n-i)g. \tag{11}$$

$E(i, j)$ and $L(i, j)$ are computed:

$$E(i,j) = e^{-c|p_e(i)-p_e(j)|}, \tag{12}$$

$$L(i,j) = e^{-c|p_l(i)-p_l(j)|}. \tag{13}$$

From these, we can compute the congruity effect:

$$C(i,j) = E(i,j) - L(i,j), \tag{14}$$

where n is the number of items in the set (list length, here $n = 26$ letters). The model had three free parameters, c, g and an overall scale factor (Table 3). Note that SIMPLE has not been developed to model response times, so we are relying on the assumption that response times and error rates are directly (linearly) related to one another. If the mapping were changed, the model output could be different. This model (Figure 10) produced a clearly curvilinear congruity effect over all probes (panel a), attenuating toward the start and end of the alphabet. Although this resembles the data from supraspan episodic lists (Liu et al., 2014), this pattern deviates from the alphabetical judgement data here. The model produced a near-null effect for probes drawn from the middle serial positions (panel c), but produced pronounced monotonic gradients for probes drawn from the first (panel b) and last (panel d) nine letters of the alphabet, deviating from the considerably flatter data. The BIC value for this model was better than all models apart from the Categorical Interference models with variable cut points (but the fit was clearly better than the Categorical Interference model with fixed cut point at the letter M).

Summary of models

Although not a perfect fit, the quantitatively best-fitting Categorical Interference model captured the qualitative features of the empirical serial-position effects the best, including the property that the size of the congruity effect tends to increase toward the start and end of the alphabet, not attenuate, as in the other models. For this to succeed, we had to assume variability in categorical cut points.

The remaining models missed key features of the data, suggesting that they may not capture the mechanism of the congruity effect for alphabetical order judgements. However, the curviness of their serial-position effects, and the attenuation of the congruity effect toward the ends of the lists do resemble congruity effects that have been previously reported, for example, in episodic-memory judgements of relative order in supraspan lists (Liu et al., 2014).

General Discussion

Testing the full range of the alphabet, the congruity effect was unequivocally found to be present in alphabetical order judgements, with response times as the measure. For error rates, at face-value, the congruity effect was not supported. However, when response times was accounted for a congruity effect did emerge for error rates, leading to the possibility that the congruity effect might act primarily to speed (versus slow) responses, with some collateral damage to accuracy via speed-accuracy tradeoffs. These results suggest that prior studies with congruity effects absent in the results may have been due to insufficient power. Thus, the benchmark finding in comparative judgements, recently extended to episodic memory judgements of relative order, has even broader boundary conditions than previously evident. This is because the alphabet deviates from typical stimuli in several ways: (a) It is long at 26 items, as compared to typical list lengths of 6–15 in both comparative judgements and episodic memory judgements of relative order. Comparative judgement studies have been conducted with larger sets, sometimes described as “infinite,” such as the classic example of animal size (e.g., Banks, 1977b). However, the alphabet is both a large, and finite (well defined) set. For such “infinite” materials, each participant might have a different conception of the exact members of the set, making it unclear how one could plot and analyze serial-position effects. The fine-structure of serial-position effects turns out to be diagnostic of model mechanisms, as we have shown here. (b) The alphabet is learned in fairly strict forward order (Zhou et al., 2006). (c) The alphabet is learned to a perfect serial-recall criterion. These latter two characteristics could have distinguished comparative judgements of the alphabet from other materials in that the forward order and perfect mastery might have neutralized the congruity effect (but they did not). (d) The alphabet has functional value (e.g., searching through dictionaries and related everyday tasks). Again, this might have set the alphabet apart from other materials.

This underlines the idea that congruity effects may, indeed, be ubiquitous within forced-choice linear continuum-based judgement tasks. However, the existing literature has suggested multiple mechanisms by which congruity effects might arise, some acting through the judgement mechanism itself, and others added on, either prior to the decision or after. Both quantitative and qualitative evaluation of our simplified models led us to favour a mechanism based on a two-level item-class effect. This account has been cast in doubt in typical comparative judgement tasks. It appears not to describe congruity effects in episodic memory tasks either. Congruity effects, although in some sense superficially similar to one another, seem to reflect quite distinct underlying mechanisms.

An important limitation of our model-selection approach is that we evaluated models only against the congruity effect. To focus on mechanisms of the congruity effect, we deliberately avoided additional complexities that would be introduced if we had modelled the full task (apart from SIMPLE, which was a full task model but still fit only to the congruity effect). If a future model could fit all the data well and the congruity effect even somewhat less well, such a model might be favoured due to parsimony and even potentially based on quantitative goodness-of-fit measures that take into account model complexity and the number of data points fit such as BIC. That said, from that perspective, our favoured mechanism of the congruity effect is separable from the judgement, itself, as was suggested by Jou et al. (2020). For alphabetical-order judgements, at least, it seems unlikely, though not impossible, that a full task model would offer different constraints on the mechanism of the congruity effect.

Viewed within the broader context of comparative judgement research, our model-selection led us to a non-parsimonious account of congruity effects. Previous comparative-judgement researchers have converged on anchor/reference-point accounts, and our own fits to memory judgements of relative order favoured directional self-terminating search for short episodic lists and a distinctiveness-based mechanism with a gradient of discriminability based on SIMPLE for longer episodic lists. Parsimony should generally be preferred unless there is strong evidence to the contrary. If a single congruity-effect mechanism were to produce slightly weaker fits of a larger set of data, a single mechanism might be favoured. The less parsimonious multiple-mechanisms position, therefore, should be further tested in future experiments. The overarching theory can be viewed as a mixture model. To counteract the large number of free parameters that a mixture model implies, one can seek principles that determine when one versus another mechanism of the congruity effect applies. For example, Liu (2015) discovered that list length could be an important factor determining whether a congruity effect arises through sequential, self-terminating search or distinctiveness. If the reason for this could be understood, list length could be integrated into a mixture model, enabling the model to more efficiently predict how the two mechanisms will trade off, given particular experimental conditions.

Given the ubiquity of stimulus-response compatibility effects (Kornblum, Hasbroucq, & Osman, 1990), it would seem rational for a participant to adjust their approach to a task to increase the odds that the probe item to which attention is first drawn is likely to be the target. This applies very directly and clearly to sequential, self-terminating search strategies such as seems to be the dominant strategy for judgements of relative order of short lists (Chan et al., 2009; Liu et al., 2014). For direct-access strategies, where position codes are compared, there is no logical advantage, as with a 2AFC procedure. If the target is not the one item, it must definitely be the other. However, with regard to human behaviour, there may also be a similar advantage for direct-access strategies. Preactivating items along a continuous positional gradient, or even categorically, could produce an overall facilitation of performance. This is particularly plausible when the instruction is predictable, as it is here, and was in the methods of Chan et al. (2009) and Liu et al. (2014).

As mentioned in the Discussion of the the experiment, simply put, the error-rate measure did not exhibit a clear congruity effect. However, there was a suggestion that a speed–accuracy tradeoff might be at play. This may at first seem circular; how could speed and accuracy trade off selectively at the end of the alphabet but not earlier on? Although clearly a post-hoc account, we propose the following. Due to the alphabet being more overlearned at earlier portions than later portions, responding too quickly would be expected to be more hazardous when the probes are later letters in the alphabet than early. Our selected model mechanism of the congruity effect is that "Late" letters are facilitated as targets for participants operating under the Later instruction, and/or "Early" letters are inhibited as targets for participants in the Later instruction (and the converse for the Earlier instruction). If, under the Later instruction, both letters happen to be Late letters, both would be facilitated (or not inhibited), biasing the participant to respond more rapidly to either probe. However, in this case, one of the "Late" probes is the correct target and the other is an incorrect response. Thus, the bias toward responding rapidly to such probes might result in fallout, increasing the error rate to some degree for probes composed of Late letters for the Later instruction. If only the participant could delay responding slightly, those errors might not be committed. Because we have assumed that memory for the earlier portion of the alphabet is superior, the effect might not be mirrored by a similar effect on Early–Early probes under the Earlier instruction. This speculation could be tested in a future study, for example, with a response-deadline. If participants are forced to respond a bit prematurely on all trials, response time would be equated, neutralizing a latency-based congruity effect, which would then presumably be relocated to the error-rate measure.

As we have discussed previously, congruity effects already require extensions to current models of order-memory. Moreover, many memory models are not developed to explain both speed and accuracy data (e.g., Brown et al., 2007; Henson, 1998; Lewandowsky & Murdock, 1989). OSCillator-based Associative Recall (OSCAR; Brown, Preece, & Hulme, 2000) is another model that has been fit successfully to JOR data. OSCAR assumes items are associated with the state of an internal context signal, and retrieving items requires re-instatement 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. Thus, OSCAR has already been used to instantiate a serial, self-terminating search mechanism, and a modification to incorporate a possible reversal of direction could produce a congruity effect resembling what has been found for short episodic lists (Liu et al., 2014). For longer lists or the alphabet, different mechanisms of congruity effects would still need to be built into OSCAR, such as a reference-point mechanism (reinstating the start-, rather than the end-of-list context), which would embody a reference point mechanism, and would thus likely be incompatible with the serial-position effects we observed.

Conclusion

Our chief result is that the congruity effect is present in alphabetical order judgements, and is best explained in terms of a categorical interference or facilitation effect, where letters are conceptualized by the participant as either “Early” or “Late.” Our findings of congruity effects in alphabetical order judgements provide evidence that the congruity effect is a general phenomenon found not only in episodic (temporally ordered) lists, but also in long semantic-memory lists that have a highly specific encoding direction. At face value, this suggests memory judgements of order may be best thought of as a subset of comparative judgements. However, comparison of numerical models to data suggests that for short episodic lists and for the alphabet, the congruity effect may result from mechanisms (direction of serial self-terminating search) that were ruled out for typical comparative judgement tasks in favour of reference-point or positional distinctiveness mechanisms, which may, in turn, apply to longer episodic memory lists. A larger repertoire of mechanisms of congruity effects must be considered, both in tasks derived from the comparative-judgement field and from the episodic-memory field, and future research will need to determine the principles and task characteristics that dictate which mechanism applies under different task demands and conditions.

Declarations

Funding

Natural Sciences and Engineering Research Council of Canada (NSERC), Alberta Ingenuity Fund.

Conflict of interest

The authors declare no conflict of interest,

Availability of data and material

Data are included in the Open Science Foundation project available at `https://osf.io/fsc8r/?view_only=983f87c245f548f9b9c8f159c0b25fdd` (Liu & Caplan, 2019).

Code availability

The MATLAB/Octave code for all models, as well as an all-purpose plotting script, are included in the Open Science Foundation project available at https://osf.io/fsc8r/?view_only=983f87c245f548f9b9c8f159c0b25fdd (Liu & Caplan, 2019).

References

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.

Banks, W. P. (1977a). 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.

Banks, W. P. (1977b). 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., Clark, H. H., & Lucy, P. (1975). The locus of the semantic congruity effect in comparative judgments. *Journal of Experimental Psychology: Human Perception & Perfromance*, *105*, 35-47.

Banks, W. P., Fujii, M., & Kayra-Stuart, F. (1976). Semantic congruity effects in comparative judgments of magnitude of digits. *Journal of Experimental Psychology: Human Perception and Performance*, *2*, 435-447.

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

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/`.

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

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.

Cantlon, J. F., & Brannon, E. M. (2005). Semantic congruity affects numerical judgments similarly in monkeys and humans. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(45), 16507-16511.

Cantlon, J. F., Platt, M. L., & Brannon, E. M. (2009). Beyond the number domain. *Trends in Cognitive Sciences*, *13*(2), 83-91.

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

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

Čech, C. G., & Shoben, E. J. (1985). Context effects in symbolic magnitude comparisons. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 299-315.

Čech, C. G., Shoben, E. J., & Love, M. (1990). Multiple congruity effects in judgments of magnitude. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(6), 1142-1152.

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.

Chen, D., Lu, H., & Holyoak, K. J. (2014). The discovery and comparison of symbolic magnitudes. *Cognitive Psychology*, *71*, 27-54.

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.

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.

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.

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

Hintzman, D. L. (2016). Is memory organized by temporal contiguity? *Memory & Cognition*, *44*(3), 365-375.

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., & Mah, W. A. (1982). Cognitive reference points in judgments of symbolic magnitude. *Cognitive Psychology*, *14*, 328-352.

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

Jou, J., Escamilla, E. E., Torres, A. U., Ortiz, A. J., & Salazar, P. (2018). Where does the congruity effect come from in memorial comparative judgments? A serial-position-based distinctiveness account. *Journal of Memory and Language*, *103*, 127-150.

Jou, J., Matos, M. S., Martinez, M. A., Sierra, F. J., Guzman, C., & Hut, A. R. (2020). Redefining the congruity effect in comparative judgments: A review of the theories and a further test. *The American Journal of Psychology*, *133*(2), 221–239.

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.

Kornblum, S., Hasbroucq, T., & Osman, A. (1990). Dimensional overlap: cognitive basis for stimulus–response compatibility— a model and taxonomy. *Psychological Review*, *97*(2), 253-270.

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

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

Liu, Y. S. (2015). *Human order memory: insights from the relative-order task* (Unpublished doctoral dissertation). University of Alberta.

Liu, Y. S., & Caplan, J. (2019). *Judgments of alphabetical order and mechanisms of congruity effects.* Retrieved 2019 Jun 21, from `osf.io/fsc8r`

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.

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

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

Marschark, M., & Paivio, A. (1979). Semantic congruity and lexical marking in symbolic comparisons: An expectancy hypothesis. *Memory & Cognition*, *7*, 175-184.

Marschark, M., & Paivio, A. (1981). Comgruity and perceptual comparison task. *Journal of Experimental Psychology: Human Perception & Performance*, *7*, 290-308.

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

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

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

Moyer, R. S., & Dumais, S. T. (1978). Mental comparison. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. 12, p. 117 - 155). Academic Press.

Murdock, B. B. (1960). The distinctiveness of stimuli. *Psychological Review*, *67*(1), 16-31.

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

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

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.

Schweickart, O., & Brown, N. R. (2013). Magnitude comparison extended: How lack of knowledge informs comparative judgments under uncertainty. *Journal of Experimental Psychology: General*, *8*(3), e54324.

Shaki, S., & Algom, D. (2002). The locus and nature of semantic congruity in symbolic comparison: Evidence from the stroop effect. *Memory & Cognition*, *30*, 3-17.

Shaki, S., Leth-Steensen, C., & Petrusic, W. M. (2006). Effects of instruction presentation mode in comparative judgments. *Memory & Cognition*, *32*, 196-206.

Shoben, E. J., Čech, 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.

Sternberg, S. (1969). Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist*, *57*(4), 421-457.

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`.

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

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.

Zhou, X., Chen, C., Zhang, H., Xue, G., Dong, Q., Jin, Z., . . . Chen, C. (2006). Neural substrates for forward and backward recitation of numbers and the alphabet: A close examination of the role of intraparietal sulcus and perisylvian areas. *Brain Research*, *1099*(1), 109-120. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0006899306001788` doi: https://doi.org/10.1016/j.brainres.2006.01.133

		Estimate (SE)
Main effects		
	Intercept	6.745 (0.015)*
	Intact/Reverse	0.007 (0.002)*
	Distance	-0.180 (0.002)*
	Instruction	-0.022 (0.022)
	Serial position (linear)	0.050 (0.002)*
	Serial position (quadratic)	-0.168 (0.001)*
	Trial	-0.042 (0.001)*
Interactions		
	Intact/Reverse × Distance	-0.046 (0.003)*
	Intact/Reverse × Instruction	0.032 (0.003)*
	Intact/Reverse × Serial position (linear)	-0.069 (0.003)*
	Instruction × Distance	-0.065 (0.003)*
	Distance × Serial position (linear)	-0.143 (0.002)*
	Distance × Trial	0.007 (0.001)*
	Instruction × Serial position (linear)	-0.136 (0.003)*
	Instruction × Trial	-0.003 (0.002)
	Serial position (linear) × Trial	0.004 (0.001)*
	Serial position (quadratic)× Distance	-0.025 (0.001)*
	Serial position (quadratic)× Trial	0.004 (0.001)*
	Intact/Reverse × Instruction × Distance	-0.021 (0.004)*
	Intact/Reverse × Serial position (linear) × Distance	-0.018 (0.002)*
	Intact/Reverse × Instruction × Serial position (linear)	-0.018 (0.004)*
	Instruction × Serial position (linear) × Distance	-0.032 (0.002)*
	Distance × Serial position (linear) × Trial	0.006 (0.001)*
	Instruction × Serial position (linear) × Trial	-0.008 (0.002)*

Table 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 * -* $p < 0.05$.

		Estimate (SE)
Main effects		
	Intercept	4.120 (0.171)*
	Intact/Reverse	0.293 (0.024)*
	Distance	-1.271 (0.026)*
	Instruction	0.256 (0.068)*
	Serial position (linear)	-0.021 (0.023)
	Serial position (quadratic)	-0.645 (0.027)*
	Response Time	-1.152 (0.025)*
Interactions		
	Intact/Reverse × Serial position (linear)	-0.448 (0.020)*
	Intact/Reverse × Distance	-0.445 (0.027)*
	Instruction × Distance	0.233 (0.026)*
	Serial position (linear) × Distance	-0.515 (0.020)*
	Serial position (quadratic)× Distance	-0.210 (0.023)*
	Instruction × Serial position (quadratic)	0.224 (0.030)*
	Instruction × Serial position (quadratic) × Distance	0.144 (0.028)*

Table 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 * -* $p < 0.05$.

Model	Free Parameter 1	Free Parameter 2	Free Parameter 3	BIC
	[range], step	[range], step	[range], step	
	fit value	fit value	fit value	
Categorical	scale			
Cut Point: M	[0.001, 2], 0.001			
	0.0662			–1839.22
Categorical	scale			
Cut Point: Full Range	[0.001, 2], 0.001			
	0.1260			**–2026.72**
Categorical	Start Letter	End Letter	scale	
Cut Point: Free	[A, Y], 1	[B, Z], 1	[0.001, 2], 0.001	
	D	W	0.100	—2021.55
Reference	log base	scale		
Point	[2, 15], 1	[0.001, 2], 0.001		
	7	0.19		–1703.14
Positional	log base	scale		
Distinctiveness	[2, 15], 1	[0.001, 2], 0.001		
	9	0.03		–1745.66
Gradient of	*c*	*g* (gradient)	scale	
Discriminability	[0.01, 0.1], 0.01	[0.005, 0.5], 0.005	[0.001, 2], 0.001	
(SIMPLE)	0.04	0.445	0.856	–1927.53

Table 3

Best-fitting models derived from direct search. Free parameters that were searched are listed along with the range and step size of the search matrix and the value of each parameter in the best-fit found for the respective model. Cells are empty when a particular model has fewer searched parameters. The lowest BIC is highlighted in boldface. Note that it is well under the $\Delta BIC = 2$ convention used for model-selection from all other models.

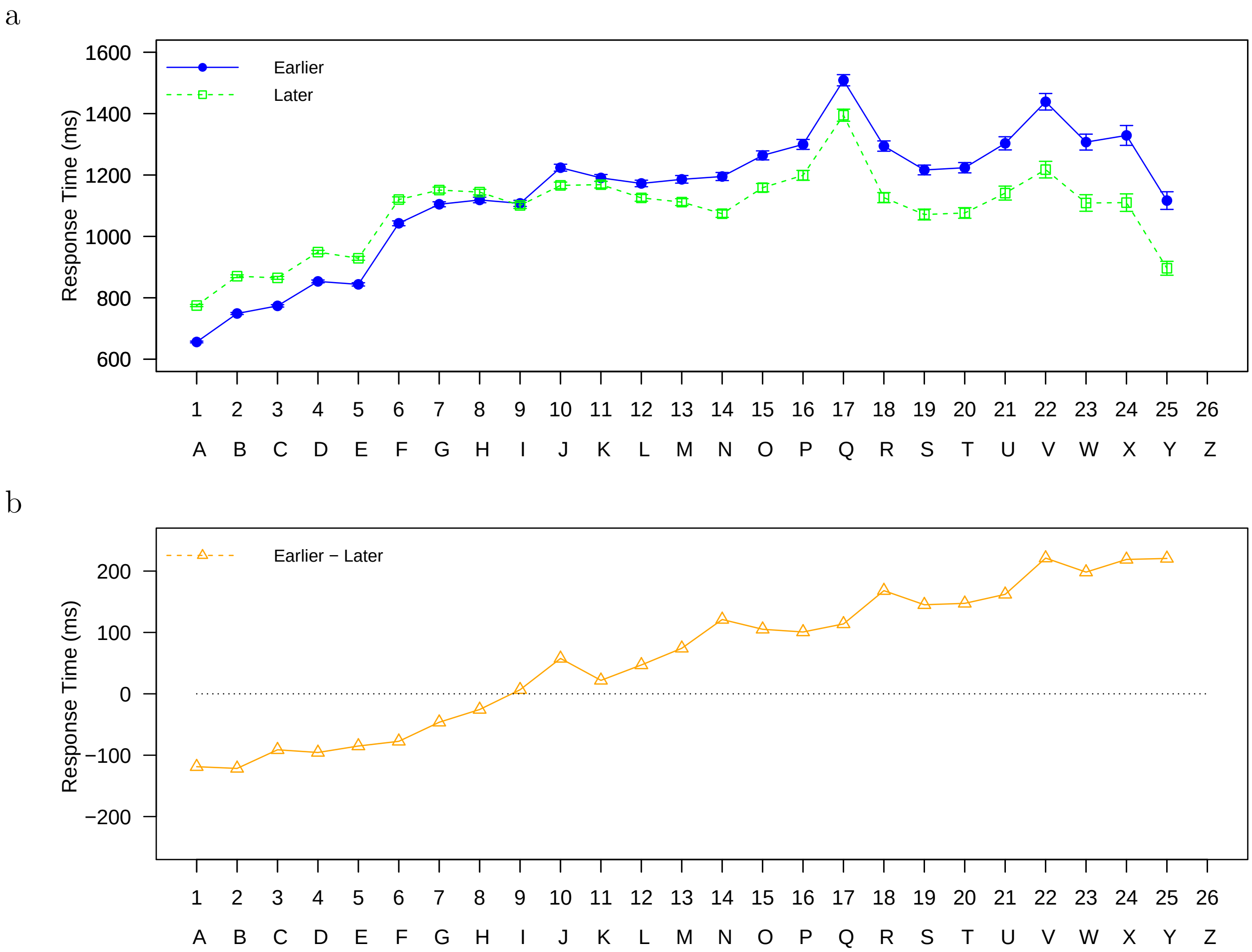

Figure 1. a) Response time as a function of Instruction and serial position. Serial position is defined as serial position of the earlier probe item. Error bars plot standard error of the mean; note that in some cases, the error bars are shorter than the height of the symbol used to plot the mean (square or circle). b) Earlier–later instruction mean differences.

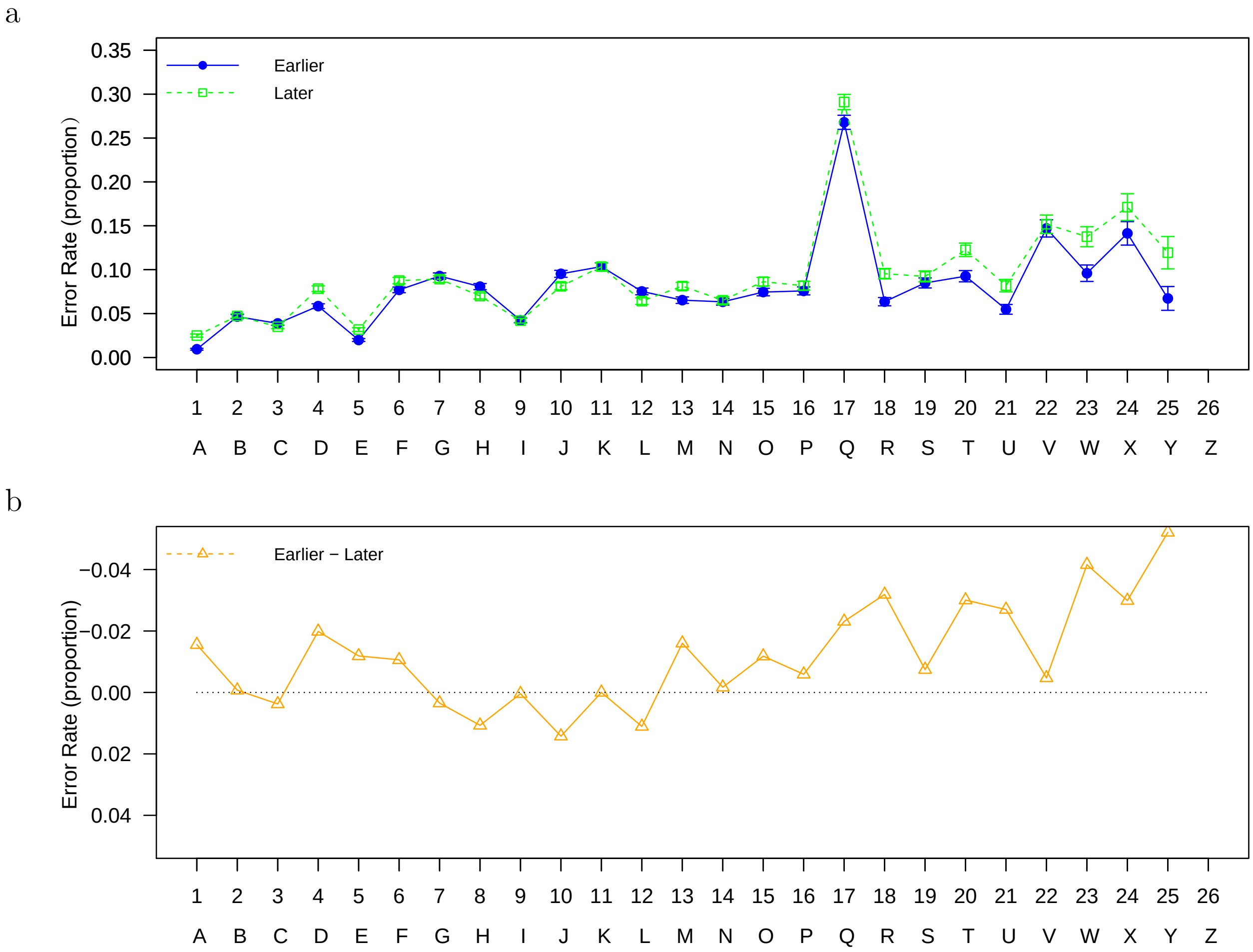

Figure 2. a) Error rate as a function of Instruction and serial position. Serial position is defined as serial position of the earlier probe item. Error bars plot standard error of the mean; note that in some cases, the error bars are shorter than the height of the symbol used to plot the mean (square or circle). b) Earlier–later instruction mean differences.

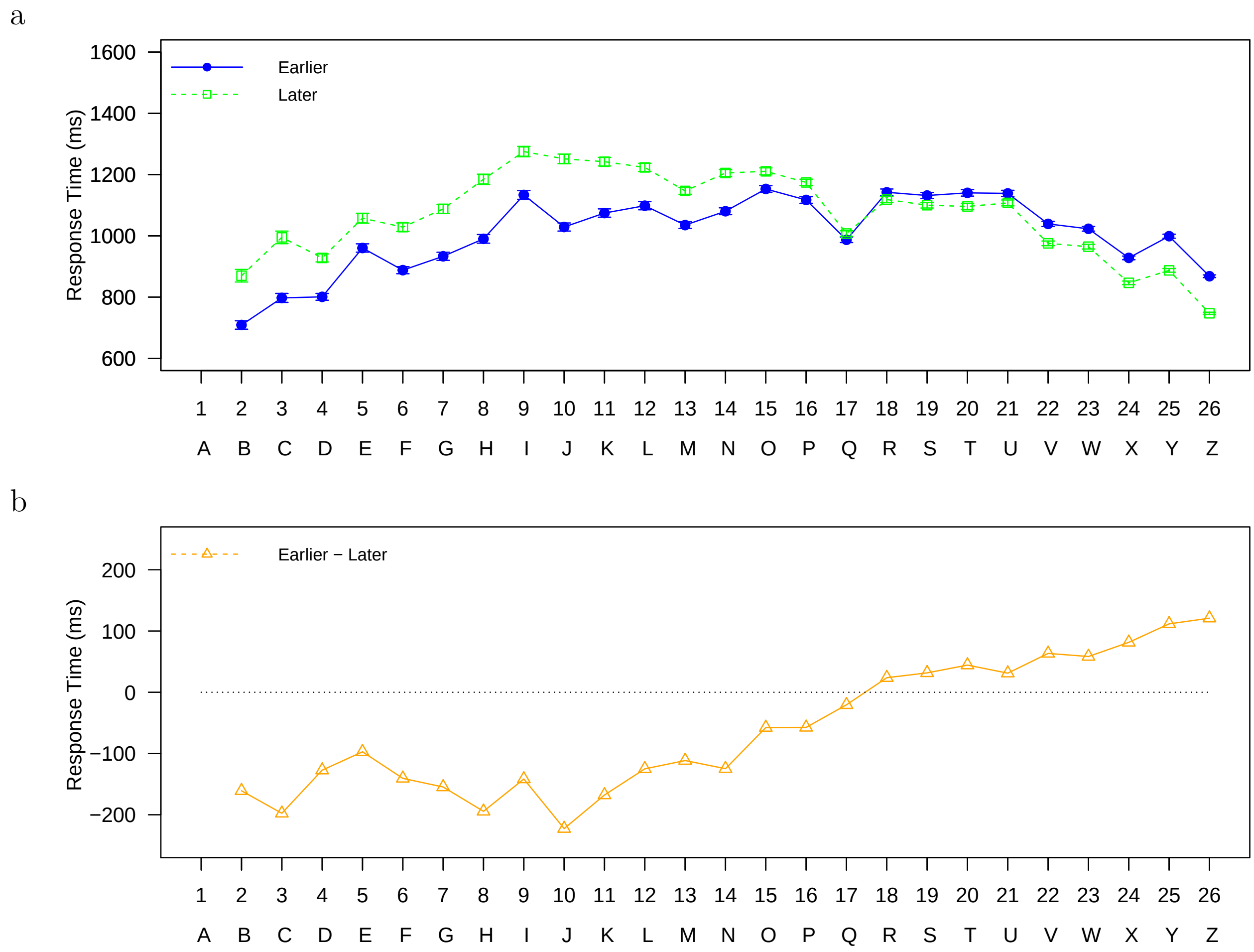

Figure 3. a) Response time and a function of Instruction and serial position. Serial position is defined as serial position of the later probe item. Error bars plot standard error of the mean; note that in some cases, the error bars are shorter than the height of the symbol used to plot the mean (square or circle). b) Earlier–later instruction mean differences.

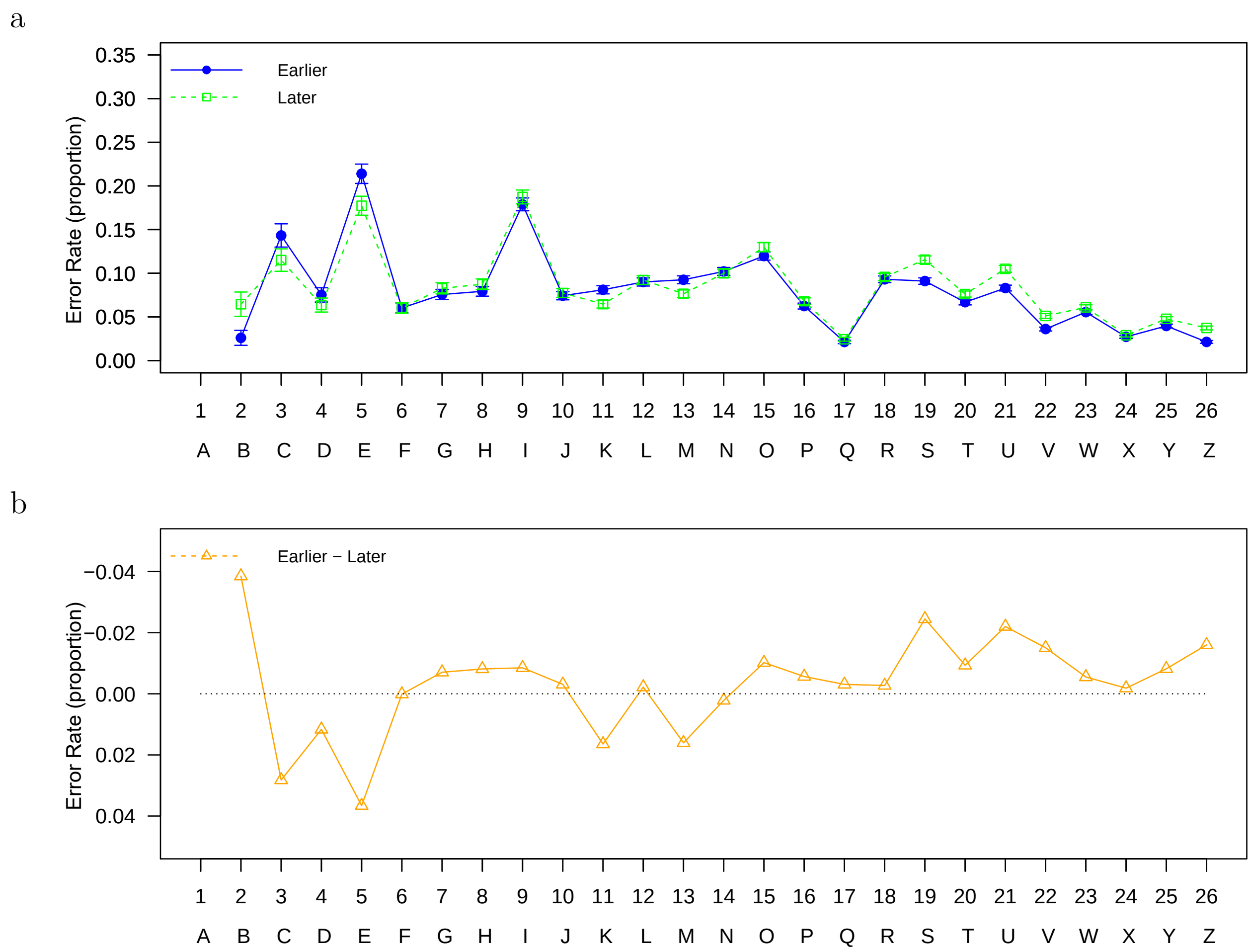

Figure 4. a) Error rate as a function of Instruction and serial position. Serial position is defined as serial position of the later probe item. Error bars plot standard error of the mean; note that in some cases, the error bars are shorter than the height of the symbol used to plot the mean (square or circle). b) Mean difference between instructions, Earlier–Later.

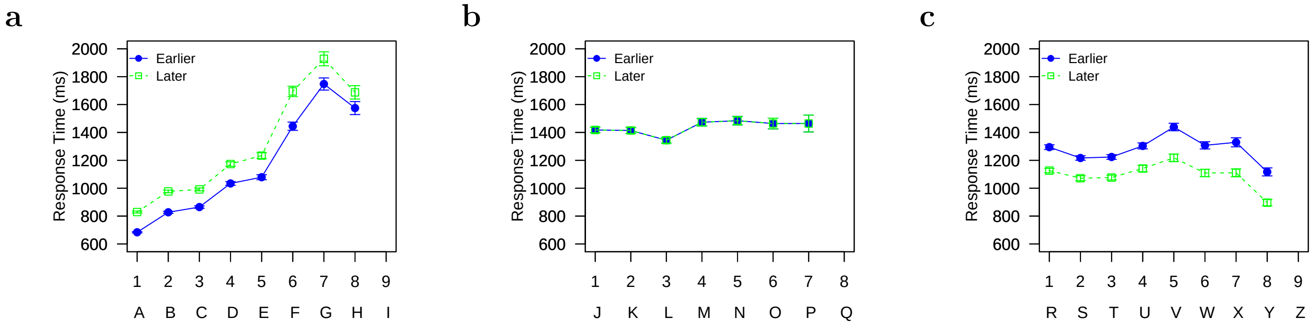

Figure 5. Response time as a function of Instruction and serial position: (a) when both probes were from the first 9 letters of the English alphabet, (b) when both probes were from the middle 8 letters of the English alphabet, and (c) when both probes were from the last 9 letters of the English alphabet. Error bars plot standard error of the mean.

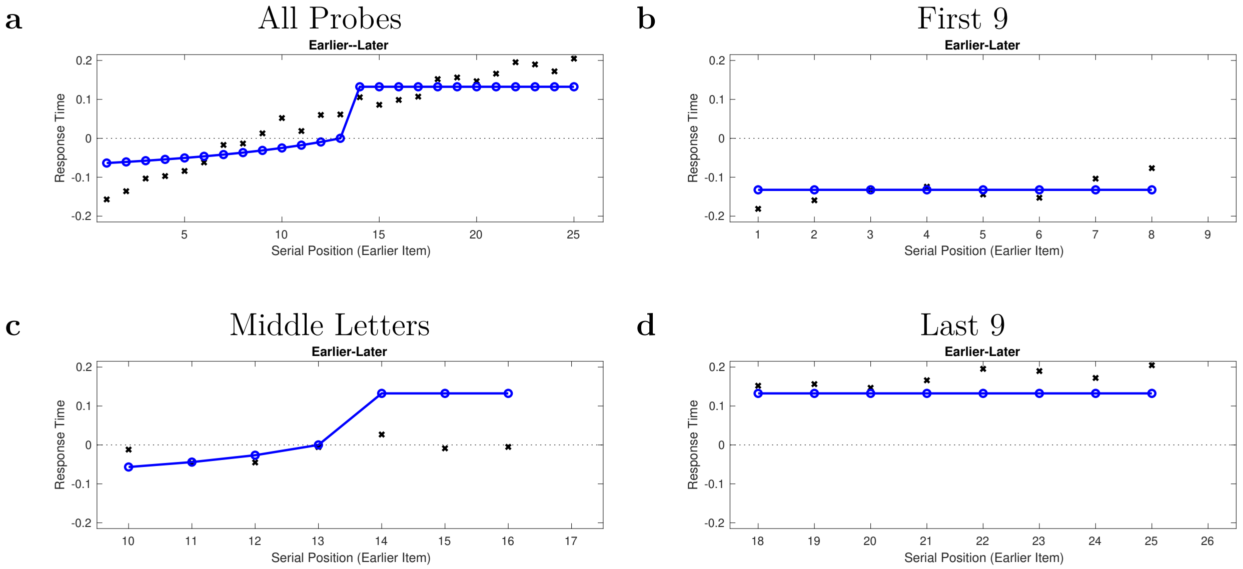

Figure 6. Model output for the best-fit of the Categorical Interference model with the early/late cut point at position 13 (letter M). Means from the empirical data are plotted with '×' and the model output is plotted in blue circles. Best-fitting parameter values and BIC are reported in Table 3.

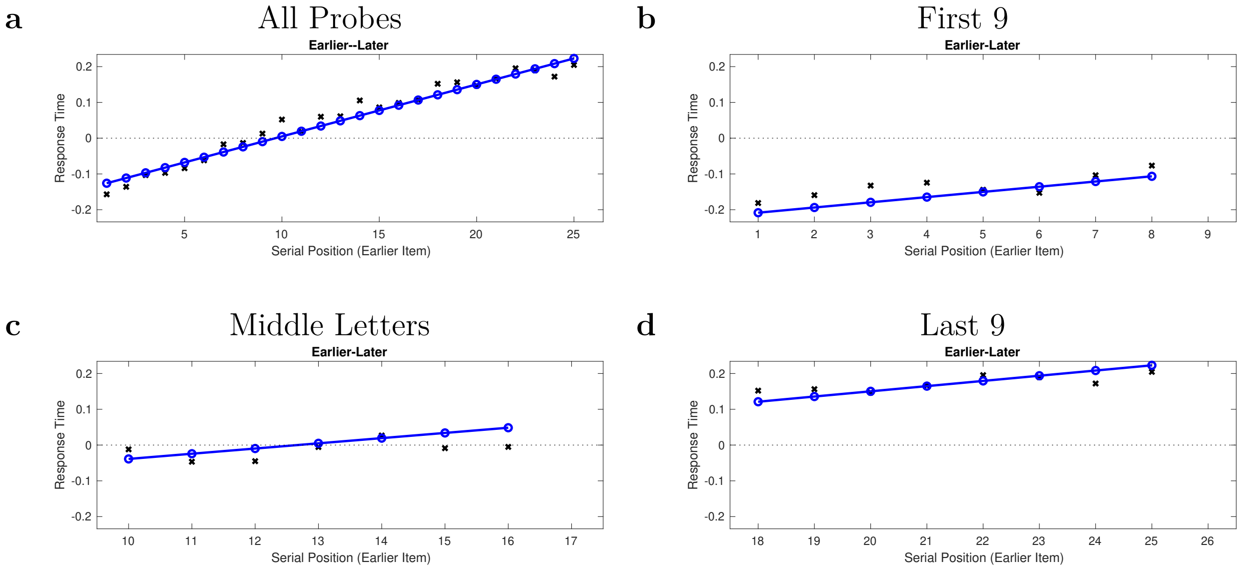

Figure 7. Model output for the best-fit of the Categorical Interference model assuming uniform variability in cut points from position 1 (letter A) to position 25 (letter Y). Means from the empirical data are plotted with '×' and the model output is plotted in blue circles. Best-fitting parameter values and BIC are reported in Table 3.

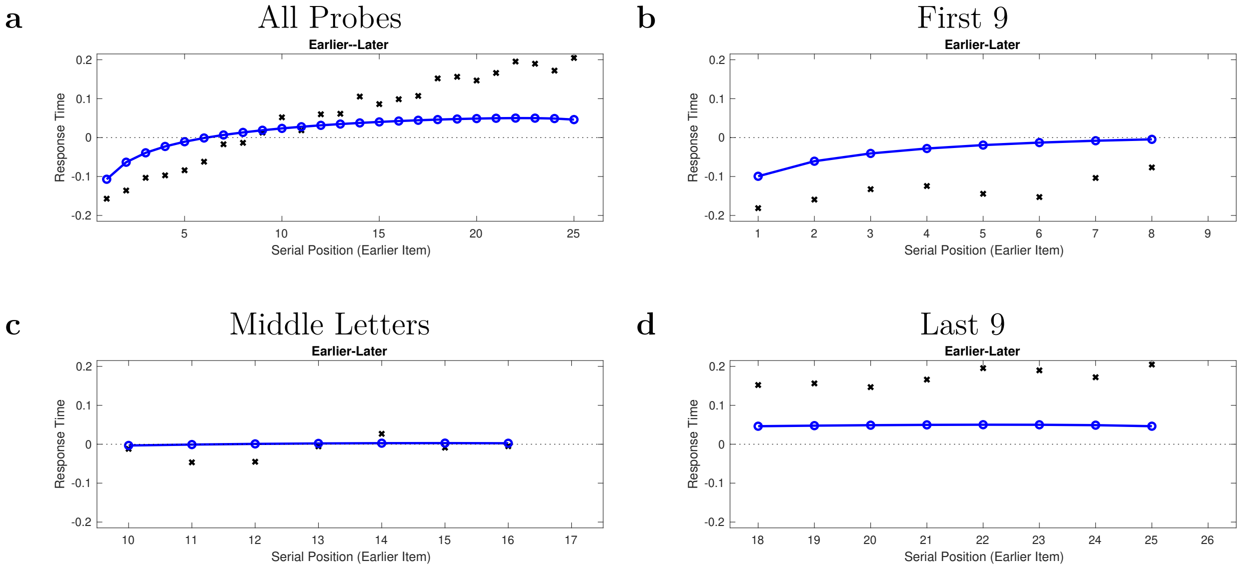

Figure 8. Model output for the best-fit of the Reference Point model. Means from the empirical data are plotted with '×' and the model output is plotted in blue circles. Best-fitting parameter values and BIC are reported in Table 3.

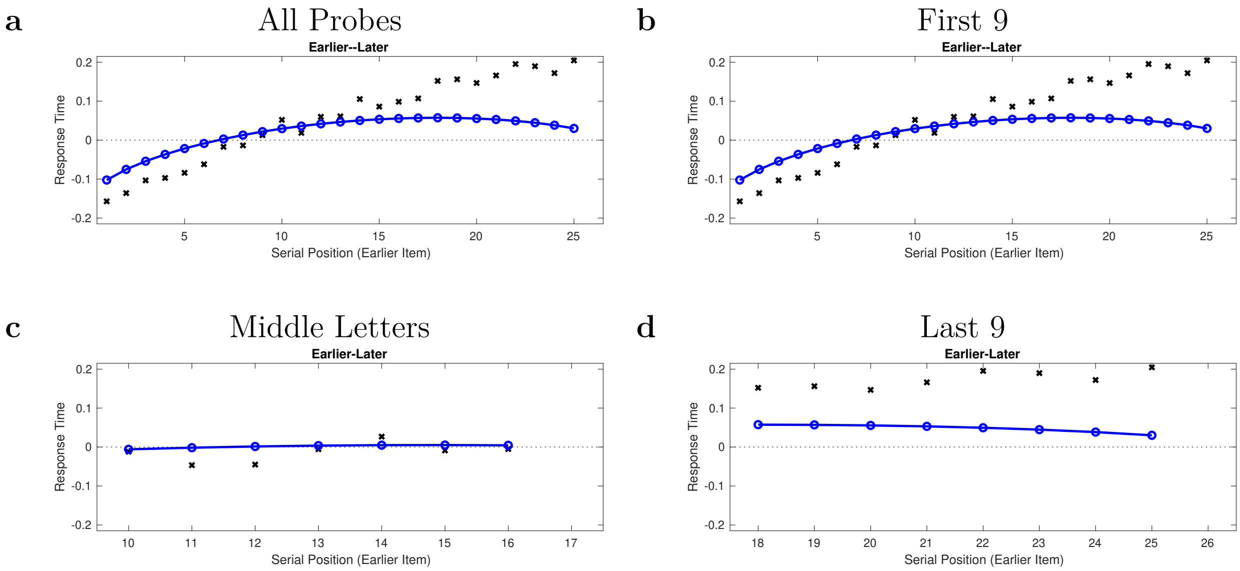

Figure 9. Model output for the best-fit of the Positional Distinctiveness model. Means from the empirical data are plotted with '×' and the model output is plotted in blue circles. Best-fitting parameter values and BIC are reported in Table 3.

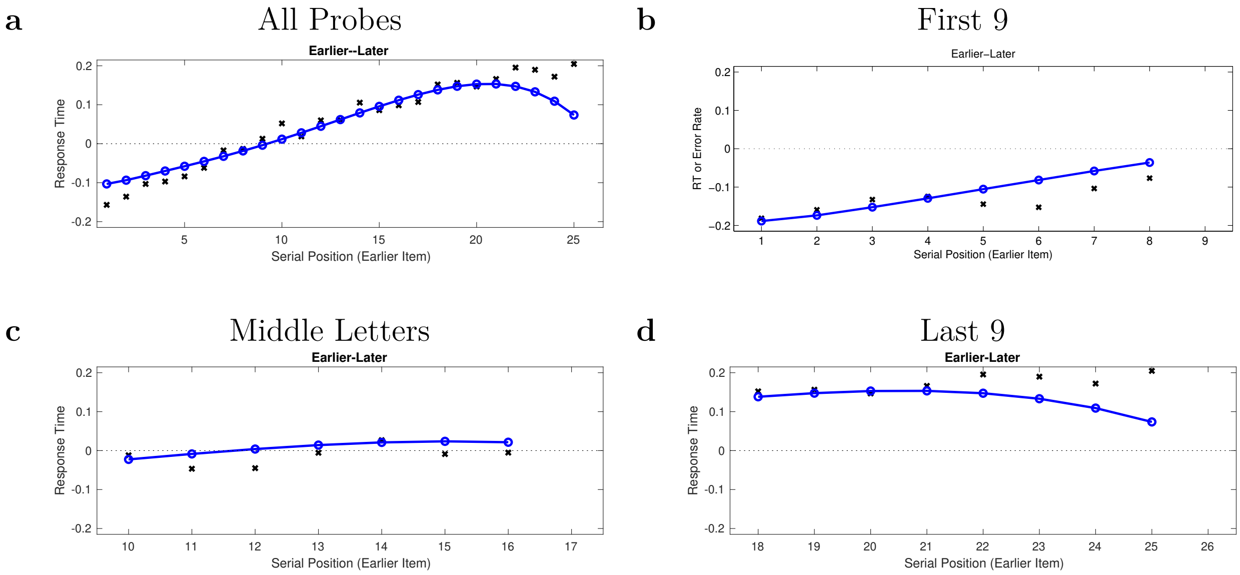

Figure 10. Model output for the best-fit of the SIMPLE-based model with a positional discriminability bias gradient model. Means from the empirical data are plotted with '×' and the model output is plotted in blue circles. Best-fitting parameter values and BIC are reported in Table 3.