**Errors in Source Memory are driven by both Systematic Intrusions from Non Targets and a Retrieval Threshold - need something a bit sexier...**The Spatiotemporal Gradient of Intrusion Errors in Continuous Outcome Source Memory: **Implications for Dual-Process Models**

Jason Zhou, Adam F. Osth, and Philip L. Smith

Melbourne School of Psychological Sciences, The University of Melbourne

Corresponding Author:

Jason Zhou

Melbourne School of Psychological Sciences

The University of Melbourne

Parkville, VIC 3052, AUSTRALIA

jasonz1@student.unimelb.edu.au

Declaration of Interest: none

Funding sources: Always better to cite current grants. It keeps the funding bodies happy. This research was supported by Australian Research Council Discovery Grant DP210101787 to Philip L. Smith and anything Adam wants listed….

Data and model code from this article can be found on our Open Science Framework (OSF) page (https://osf.io/76wtm/). This experiment was not pre-registered.

Abstract

Previous research has characterized source retrieval as a thresholded process, which fails on a proportion of trials and leads to guessing , as opposed to a continuous process, where response precision varies across trials but is never zero. The thresholded view of source retrieval is largely based on the observation of heavy tailed distributions of response errors, thought to reflect a large proportion of “memory-less” trials. In this study, we we investigate whether these errors might reflect systematic intrusions rather than source guessing. Intrusions occur when information about a nontarget item is erroneously reported instead of the target item and have been observed in tasks ranging from serial recall to visual working memory. When the studied items are uniformly distributed , then intrusion from nontargets at test may appear uniform, implying responses in the absence of information, despite actually being driven by information for a different item. Using the circular diffusion of decision making we found that intrusions account for some, but not all, errors in a continuous-report source memory task. Additionally, we found that a model in which the probability of a given nontarget intruding is determined by the spatiotemporal similarity of its study event and that of the target provided the best account of response error and time data. Our findings support a thresholded view of source retrieval, but suggest that previous work has overestimated the proportion of guesses which have been conflated with intrusions.

*Keywords:* source memory, intrusion, swap error, contiguity, response times

When we recall a past experience, we often not only retrieve information about an item in memory, but also information about the conditions under which that memory was formed, or the *source* of that memory (Johnson et al., 1993). Episodic memory, which describes memory for events, has been studied experimentally using item recognition and source memory tasks, often in tandem. In contrast to recognition tasks, where the focus is on the presence or absence of an item in the study episode, the focus in the source memory task is on the particular context in which items are studied. In a typical source memory task, subjects are shown stimuli (e.g., words, shapes, or objects) which are presented in some context (e.g., the voice of a speaker, location on a display). When later cued with the item, participants are then asked to report the source. In these types of tasks, the source of an item is a part of the associative information that is encoded in memory, which can act as a cue for retrieval. Source memory tasks are theoretically important because they offer insight into how items become associated with contexts, which bears on how information is organized and stored in memory. Several models have been advanced to understand the processes governing both recognition and source judgements (e.g., Yonelinas, 1999; Slotnick & Dodson, 2005; Hautus et al., 2008).  
 A key theoretical questions these models seek to answer is whether the retrieval of information from source memory is better characterized as a continuous or a discrete process. In continuous models of source memory, which are based on Signal Detection Theory, memory strength is assumed to vary continuously, and so predict that performance in a source memory task declines gradually as memory strength decreases (Banks, 2000; Mickes et al., 2009). In contrast, threshold or discrete-state models assume that memory strength for an item must reach a certain threshold in order for that item to be retrieved, and so predict that source responses are either made with high precision when driven by memory or are guesses, made in the absence of information, when the memory is below the retrieval threshold (Batchelder & Riefer, 1990; Klauer & Kellen, 2010). These alternatives are not mutually exclusive but may co-exist, which leads to dual-process models. These models propose that different retrieval mechanisms are used in different kinds of memory tasks (Mandler, 1980). In the influential Yonelinas (1999) dual-process model the two processes are 1) familiarity, which yields a continuous measure of strength for an item in memory and 2) recollection, which yields rich information about the study event itself when memory strength exceeds a threshold, but fails absolutely below that threshold. In a recognition task, responses can be made either by directly retrieving an item from memory via recollection or by making a judgment about whether it is in memory without retrieval based on a feeling of familiarity. In this way, both recollection and familiarity can contribute to successful recognition. In a source memory task, however, familiarity cannot distinguish between two studied items from different sources, because both items are present in memory and should therefore be equally familiar. Thus, the Yonelinas (1999) dual-process model predicts that source judgements should rely purely on a high threshold recollection process.

The dual-process account of recollection only holds if source memory retrieval is actually a thresholded process. Research which has attempted to distinguish between continuous and thresholded models of source memory has been largely based on rating-scale data from two-choice tasks and used confidence ratings to construct Receiver Operating Characteristic (ROC) curves (Yonelinas, 1999; Slotnick & Dodson, 2005). Although the predicted shape of these curves were initially thought to distinguish between continuous and thresholded models, subsequent work has found numerous conditions under which the models mimic each other (Yonelinas & Parks, 2007; Klauer & Kellen, 2010). [I think the structure of the Introduction is good and includes the right amount of detail to motivate your study without belaboring the point unnecessarily.]

**Continuous-Outcome Tasks**

In response to the model-mimicry problem, researchers have turned to tasks that provide richer data than are obtained from traditional signal detection tasks in an attempt to try to distinguish the models. One such alternative, often used in the study of visual working memory (VWM), is the continuous-outcome task, in which responses are made on a continuous scale (Wilken & Ma, 2004). Two widely-studied tasks in the VWM literature involve memory for color and orientation. Participants are asked to reproduce the color or orientation of studied items by selecting corresponding points on a color wheel or response circle (Zhang & Luck, 2008; van den Berg et al., 2014; Adam et al., 2017; Smith et al., 2020). In the present study, participants are shown words positioned continuously around the perimeter of a circle, and then later asked to reproduce the location of the cued word. The advantage of using such a task is that it allows direct measurement of response precision, which characterizes the magnitude of the response error, as opposed to the proportion of responses in each of the discrete options in a traditional two-choice task, which simply characterizes whether the response was correct or incorrect. This richer, continuous measurement is more informative about the nature of mental representations, particularly in terms of the variability of decisions made about these representations (Smith et al., 2020).

Just as the source memory literature has been concerned with the question of retrieval thresholds, the VWM literature has historically grappled with whether storage capacity is determined by a discrete number of “slots” to be filled, or a continuous resource that can be distributed across an increasing number of items that are represented with decreasing resolution in memory. Once again, these alternatives are not mutually exclusive: memory may be both item-capacity limited and resource limited (Donkin et al., 2016; Sewell et al., 2014; Zhang & Luck, 2008). In these kinds of hybrid slots-resources models, the precision with which items are represented in memory depends on the resources allocated to them within an overall item-capacity limit. Retrieval may fail because an item is not in memory, in which case the participant guesses, or because it is in memory but that the resources allocated to it leads to it being represented with low precision. [Something like perhaps?]

In both cases, the common question about the architecture of memory is if information is stored in discrete states. Zhang and Luck (2008) modelled distributions of response outcomes in a color recall task under different set size conditions, and found the data was well described by a mixture model, specifically a mixture of a von Mises distribution[[1]](#footnote-2) and a uniform distribution, which they interpreted as reflecting a combination of high-accuracy memory-based decisions and guessing, supporting the slots model of VWM capacity in the same way that thresholded views of source memory claim that retrieval is “all-or-none”: information is either stored with high resolution in memory it isn’t present at all. Bays et al. (2009) proposed an extension of the model in which the precision of the item representations varies with the number of items in memory and in which, on some trials, the wrong item is reported. This model is related to the intrusions model we present below. Van den Berg et al. (2014) proposed a variable-precision model in which both the number of items in memory and the precision with which they are represented varies from trial to trial. Variable precision is also a feature of the intrusions model we present below. [This connects to the point I was making about about different reasons for retrieval failure.]

Harlow and Donaldson (2013) used a similar approach to study source memory. They used a continuous-outcome task in which word stimuli were paired with locations on the circumference of a circle, which were defined as the “source”. At test, participants were cued with words and were required to remember the source location by moving a mouse to the corresponding point on the response circle. The authors found that a mixture model consisting of a wrapped Cauchy and a uniform component was preferred over a pure wrapped Cauchy model, which was interpreted as evidence for a thresholded retrieval process which yields uniform guesses when memory strength is subthreshold (Harlow & Donaldson, 2013). The Cauchy distribution, like the normal distribution, has a bell-shaped probability density function, but unlike the normal distribution, its variance is not finite. Harlow and Donaldson (2013) chose to use a wrapped Cauchy distribution instead of a wrapped normal because of its heavier tails, which better characterized the distributions of errors in their continuous-outcome task than a normal distribution, but nevertheless found the empirical distribution was better described by a model that combined the Cauchy and uniform distributions. Unlike the normal distribution, which is assumed on theoretical grounds in signal detection theory, the Cauchy distribution in their analysis was not intended to be a model of the retrieval process but was simply an empirical model of the distribution of errors in the data.

The Harlow and Donaldson (2013) interpretation attributes variability in response precision to two processes : 1) variability in the precision of items in memory and 2) the possibility that the item is not in memory and the response is a guess. To account for the contribution of decision processes to response variability, Zhou et al. (2021) applied the circular diffusion model (Smith, 2016) to a source memory task using Harlow and Donaldson’s (2013) paradigm. Unlike empirical characterizations of response error, like the one provided by the wrapped Cauchy model, the predicted distribution of response errors in the circular diffusion model is derived from an evidence accumulation model of the retrieval process. Also unlike the wrapped Cauchy model, and similar models used to characterize performance in the VWM literature, the circular diffusion model predicts both distributions of retrieval errors and distributions of response times (RT). The latter play an important role in the study we describe below.

### Decision-Making in Continuous-Outcome Tasks

Any response observed in a memory task is a product of a decision process acting on the representation of stimulus information retrieved from memory. Accurately characterizing the effect of decision-making is critical to understanding the nature of memory retrieval (Ratcliff, 1978). The importance of modeling decision-making is well illustrated in the recognition memory literature. As mentioned previously, much of the past work characterizing recognition memory has been based on ROC shapes. In particular, DeCarlo (2002, 2003) and Yonelinas (1994) used curvilinear z-transformed ROC (z-ROC) functions, which are expressions of response proportions at different levels of confidence, as evidence for a mixture of two continuous processes or a mixture of a continuous and discrete processes respectively. Ratcliff and Starns (2009) noted that to analyze response proportions exclusively is to overlook another major dependent variable in RTs. By modelling the process by which different confidence judgements compete to reach different evidence accumulation criteria, Ratcliff and Starns (2009) demonstrated that a variety of z-ROC shapes could be produced with changes in these decision criteria, with a single underlying memory process, invalidating the previous theoretical conclusions based on response outcomes alone (also see Starns et al., 2012; Dube et al., 2013).

In another example from the free recall literature, Osth and Farrell (2019) compared different mechanisms that aimed to explain the advantage for items at the start of the list (the primacy effect). In a demonstration of the diagnostic power of full RT distributions, Osth and Farrell (2019) found that while strength, rehearsal, and reinstatement based explanations all made similar predictions about the shape of serial position curves and mean RTs, they made distinct predictions about how the shape of RT distributions changed across serial positions, which revealed a consistent preference for the reinstatement account when modeled with the diffusion decision model, which we describe below.

The diffusion decision model is a particularly influential account of decision making, which successfully explains well-documented phenomena like the speed-accuracy trade-off, and slow and fast error patterns under different decision conditions (Ratcliff et al., 2016). The diffusion model describes decision-making as a noisy evidence accumulation process, the rate of which is defined as the *drift rate*, that accumulates until a response boundary or criterion that represents the amount of evidence required for a given response to be output (Ratcliff & McKoon, 2008). Variation in decision criteria can reflect response bias, for example, decision-making under speed emphasis can be represented with a lower criterion relative to emphasizing accuracy. Drift rate reflects the quality of evidence driving the decision process and draws an explicit link between response accuracy and RT: higher drift rates result in higher accuracy and faster RTs, while lower drift rates result in lower accuracy and slower RTs (Ratcliff et al., 2015). In applications of the model to memory, the drift rate reflects the quality of the information retrieved from memory, estimates of drift rate from the model and the way in which they varying across experimental conditions are important theoretically in testing between alternative models of the memory system.

The circular diffusion model (Smith, 2016) is an extension of the diffusion decision model to model continuous decision outcomes and inherits the desirable explanatory qualities of the standard two-choice diffusion model. Decision-making is represented as evidence accumulation in two-dimensional space that begins at the origin of a circle and terminates at a point in its circumference, which represents the outcome of the decision. Because the diffusion process is two-dimensional, the drift rate is defined as a vector with a direction, or  *polar angle*, (I’m starting to prefer this as less jargony) that represents the identity of the encoded stimulus, and a length or *norm*, which represents the quality of the encoded stimulus (Figure 2). The norm of the drift vector determines the RT in the same way that scalar drift rate does in the Ratcliff (1978) model.

Figure 1

*Circular Diffusion Model of Continuous Report*

A close up of a device

Description automatically generated

*Note.* Evidence accumulation is represented by a 2D diffusion process on the interior of a disk whose bound circle, of radius *a* is the decision criterion for the task. Evidence accumulation begins at the center of the disk and continues until the process hits a point on the bounding circle. The hitting point, Xθ, is the decision outcome and the hitting time, Ta,is the decision time. The drift rate is vector-valued and consists of two components, (μ1, μ2), which jointly specify its magnitude and direction. In polar coordinates the magnitude is represented by the drift norm ||μ|| and direction is represented by the angle θμ The noisy sample path represents evidence accumulation on a single experimental trial. From P. L. Smith (2016). “Diffusion theory of decision making in continuous report’ Psychological review, 123, 425-451. Figure 2. Copyright American Psychological Association.

When the drift rate and the decision criterion are fixed across trials, the circular diffusion model predicts that the decision outcomes follow a von Mises distribution. The dispersion of outcomes in the von Mises distribution depends on a precision or concentration parameter, κ, which is jointly a function of the drift norm, ||μ||, the decision criterion, *a*, and the noise in the evidence accumulation process, σ2:

|  |  |
| --- | --- |
|  | (1) |

which defines a clear relationship between the strength of evidence and decision criterion in determining the observed distribution of responses (Smith, 2016).

Through across-trial variability decision-making, specifically drift variability in the circular diffusion model, a single continuous process can produce distributions of response error with heavy-tails through the decision-making process, without invoking mixture with a uniform component in the memory process (van den Berg et al., 2012; Smith, 2016). Across-trial variability in drift rate is the circular diffusion model’s counterpart of variable precision, as assumed in the successful model of visual working memory of van den Berg et al. (2014). In a previous study, Zhou et al. (2021) investigated whether this property of the diffusion decision model could account for the distribution of errors in source memory retrieval observed by Harlow and Donaldson (2013), without needing a threshold in the memory retrieval process. To do so, we (is “we” referring to the previous study or the current one? It’s ambiguous syntactically. Better as “they”?) compared three different variants of the circular diffusion model: 1) a single diffusion process with across-trial drift rate variability, 2) a two-component mixture of a diffusion process with drift rate fixed across trials, and a zero-drift process (i.e. decision-making in the absence of evidence), which represented guessing for a proportion of trials where memory strength was subthreshold and 3) a hybrid two-component mixture model with across-trial variability in drift rates for the positive drift process. We found that the latter two models, with the zero-drift mixture, provided a consistently better account of response error and response time (RT) data than the model with drift rate variability but no zero-drift guessing. Furthermore, we found that although drift rate variability in the hybrid model resulted in marginal improvements to fit relative to the thresholded model, when each model was penalized for the number of freely estimated parameters, the thresholded model was preferred as the more parsimonious model. Our results showed that the heavy-tailed distribution of errors was not attributable to variability in decision-making (a decision phenomenon), but instead evidence of guessing on a proportion of trials (a memory phenomenon), corroborating the initial conclusions of Harlow and Donaldson (2013) and the thresholded view of source memory retrieval.

[The material reads as a bit weak as the primary motivation for the study. Specifically, it reads like a secondary point rather than its main purpose. I’d lead with your strong suit and then introduce the methodological point later. Something like: “As we noted previously, the VWM literature has highlighted the importance of distinguishing errors due to intrusions from errors due to guessing. Bays and colleagues, in particular, have shown that heavy-tailed distributions of errors, which the Zhang and Luck (2008) slots-plus-resources model attribute to guessing, can be predicted by a pure resources model in which the heavy tails are due to intrusions from other items in the display. They argue that if intrusions are not taken into account then estimates of the guess rate will be inflated, which will lead to the evidence for slot models being overestimated. Our aim in this study was to investigate the effects of intrusions on source memory retrieval. Our motivation for doing so was like that of Bays and colleagues: If intrusions from other items in a list contribute significantly to retrieval errors and if the effects of intrusions are not characterized explicitly, then their effects will be misattributed to guessing. In the context of the source memory debate, this would lead to an overestimation of the evidence for a thresholded retrieval process. The primary theoretical contribution of our study is to present and evaluate a spatiotemporal gradient model of the intrusion process. This model shows that intrusions from other list items do indeed contribute significantly to the distribution of retrieval errors and that the magnitude of the intrusions depend both on the temporal proximity of the target and distractor items in the list and the spatial proximity of their source locations. By characterizing intrusions systematically in this way we are able to distinguish their effects from those of guessing. This allows us to provide a more principled evaluation of whether or not retrieval is thresholded.” [Most of the detailed stuff about gradient effects you review below is fine, but I would lead with this material, and provide some kind of summary statement like the one above to tell people what’s new and important about this study.]

A secondary aim of our study was methodological and concerned the ease with which an association can be formed between an item and a source location and then subsequently retrieved. This question is important for the evaluation of the Harlow and Donaldson (2013) and Zhou et al. (2021) studies, both of which defined the memory source as a designated point on the circumference of a circle that was associated at encoding with a centrally presented target word. With this mode of presentation the association between the item and its source is implicit rather than explicit and requires participants to form the association at stimulus encoding. Consequently, it may lead to the difficulty in source retrieval being overestimated. A secondary aim of our study was to compare the Harlow and Donaldson and Zhou et al. mode of presentation to one in which the items were presented at specific locations on the response circle, which made the association between item and source explicit.

….Then your other stuff below...

In the current study, we follow up on the findings of Zhou et al. (2021), by considering two points that may challenge previous thresholded interpretations of the continuous-outcome source memory data. The first of these points relates to the experiment design used by Harlow and Donaldson (2013) and inherited by Zhou et al. (2021), specifically regarding how stimuli were presented. Instead of simply presenting words in locations, source locations were represented by crosses located along the circumference of a circle, which were then replaced with the presentation of the word in the center of the screen. The temporal and spatial dissociation between item and source information caused by presenting word/location pairs in this sequential manner may have contributed to participants’ inability to accurately recall source locations when cued with words at test. Although participants were asked to verify the hidden location after studying words to ensure both components were encoded, it is possible that the appearance of source guessing is attributable to losses of bindings between a proportion of word/location pairs after successful encoding of the separately presented components, rather than a retrieval threshold in the memory process. Instead, if source retrieval is actually thresholded, then the observation of heavy-tailed source errors should be robust to changes in how the stimuli are presented. To address this point, we explicitly manipulate the presentation format of the stimuli across participants, comparing a condition in which word/location pairs are presented sequentially with a second condition in which words in the location they are to be associated with (we detail this manipulation in the method section to follow). Secondly, while heavy-tailed distributions are often interpreted as evidence for a uniform distribution of guesses, other memory processes may give rise to apparently uniform errors without needing a separate guessing process. In the section to follow, we consider the possibility that participants sometimes confuse target and non target items, which would give rise to apparently uniform errors when all items are uniformly distributed in the report feature space, as they are in the Harlow and Donaldson (2013) design.

## Non target Responding

I think this can be handled better by putting it into the motivating preamble in the way I have suggested. That way you link your gradient model explicitly to the larger thresholded retrieval debate.

In the VWM literature, the slots account of memory capacity proposed by Zhang and Luck (2008) is built upon the finding that a proportion of responses appear to be uniformly distributed and reflect random guessing. Bays et al. (2009) challenged this interpretation by arguing that confusions between target and non target items could also account for errors that appear uniform relative to the target item. Figure 2 illustrates an example of how intrusions drawn from distributions centered on randomly dispersed non targets can mimic guesses from a uniform distribution.

Figure 2

*Different Sources of Error in the Continuous-Outcome Task*

Shape, circle

Description automatically generated

The tendency for subjects to respond to non target features or items has been observed in a wide variety of cognitive tasks, and the related types of errors that arise are referred to by various terms including binding, transposition, intrusion, and swap errors[[2]](#footnote-3), each reflecting specific properties of the tasks used to study the phenomenon (Bays, 2016). Confusions between items can be confounded for variability in memory precision, and so disentangling these sources of error has been important in accurately characterizing VWM processes (Bays, 2016; Rajsic & Wilson, 2012, 2014; Pertzov et al., 2012). Because the distribution of target features (in this instance, color) are random, response errors arising from confusions between a target and non target item will also appear to be uniformly distributed. Consequently, non target responses of this kind are indistinguishable from guessing when measured from the target. However, Bays et al. (2009) noted that the two sources of error are easily differentiated by measuring the frequency of responses relative to all non targets: guesses are uncorrelated with non target items, so the Zhang and Luck (2008) predicts that that the resultant distribution should be uniform. Instead, the authors found clear evidence of central tendency in the distribution of responses relative to non targets (Bays et al., 2009). That responses centered on non target items are more frequent than expected by chance is interpretable as evidence for non target responding. Furthermore, when these non target responses were accounted for, model estimates for the proportion of guesses decreased dramatically and no longer required an upper limit to the number of stored items. In the current study, we take a similar approach to Bays et al. (2009) and consider whether prior estimates of the rate at which people guess when making source memory judgements hold when the possibility of non target responses are accounted for.

One challenge in distinguishing between errors arising due to random guesses and swaps is that different model assumptions can result in different estimations of swap rates in VWM tasks (Williams et al., in press). In the present study, we seek to address this challenge by using a richer data space, specifically by jointly modelling RT and error data using the circular diffusion model. In addition, we compare models which make different assumptions about the effect of similarity between items on the rate of intrusions in the source memory task.

I’ve suggested moving it right up to the beginning...

Most explanations of non target responding attribute the phenomenon to confusion between items that are similar (Rerko et al., 2014; Bays, 2016; Oberauer & Lin, 2017; but see Pratte, 2018 for an alternative view). To extend the reasoning of Bays et al. (2009), if intrusions from non targets are driven by confusions between items, then the probability of a given non target item intruding should systematically vary with the degree of similarity between that item and the target. In the continuous-outcome source memory paradigm, items may be similar in several ways, including the position of items in the study list, the spatial proximity of the item sources, as well as in the semantic and orthographic features of the words themselves. Many of these effects have been studied in the larger episodic memory literature. In the section to follow, we review the commonalities between findings across different memory tasks, all of which motivate the present modeling of intrusions in source memory.

### Contiguity Effects

The principle of *temporal contiguity* is that events that occur close in time become associated with each other (for an extensive review of contiguity effects in episodic memory, see Healey et al., 2018). One interesting way in which temporal contiguity has been studied is using free recall tasks, in which participants are asked to recall a list of items in any sequence they wish. Participants’ responses in free recall tasks are interesting because they are illustrative of how items are spontaneously organized in memory (Howard & Kahana, 2002). In particular, Kahana (1996) demonstrated that after recalling a given item, the next item to be recalled tends to be a neighboring item in the study sequence. The distance between item *i* in the study list and another item in the list is known as the *lag*. Neighbors in the forwards direction (*i* + lag) were more found to be likely to follow an item than backwards neighbors (*i* – lag), referred to as forward asymmetry.

While associations between temporally contiguous items can facilitate responses in free-recall tasks, the same type of association can contribute to errors in tasks when the sequence of items is important. Specifically, in serial recall tasks, when subjects must recall lists of items in the sequence in which they are given, a classic finding is that incorrect responses tend to be items studied near the target in the study sequence, which can be described as a transposition error in the output list (Lee & Estes, 1977; Nairne, 1990). As with free recall, serial recall data forms a transposition gradient around the target location effect (Kahana & Caplan, 2002; Solway et al., 2012, which tends to exhibit a forwards asymmetry both in terms of transposition probability (Klein et al., 2005; Haberlandt et al., 2005[[3]](#footnote-4)) as well as latency (Farrell & Lewandowsky, 2004; Hurlstone & Hitch, 2014). The temporal contiguity effect has also been observed in paired-associate recall. After studying pairs of words, Davis et al. (2008) found that when participants recalled non target items, the erroneous item tended to be intrusions from temporally contiguous pairs. Again, the probability of an intrusion from pair *i* + lag when cued with an item from pair *i* decreases with absolute lag, and is asymmetric in the forwards direction so that intrusion probability is greatest when lag = 1.

Taken together, these list-learning paradigms demonstrate that participants are sensitive to the temporal context in which items are studied and this has a strong influence on how these items are represented and retrieved from memory under a variety of different task demands. As such, precise characterization of source memory retrieval requires an account of the effect of temporal contiguity (and other forms of contiguity explored in the following subsections) on source retrieval. Recently in the source memory literature, Popov et al. (2021) investigated errors in binding between words and the locations along a circle in which they were presented, and found that when participants made a misbinding error, responses were not generated from a random non target. By separately estimating the probability that a response came from each of the locations in the study set, Popov et al. (2021) found that mis-binding errors were most likely to come from locations in neighboring serial positions, demonstrating a relationship between the probability of binding errors and serial position that can be explained as an effect of temporal contiguity. Building upon this finding, in our present modelling (described formally later), instead of freely estimating the probability of intrusions from each lag, we constrain the effect of temporal similarity our model to make more systematic predictions about the relationship between the two, specifically predicting that intrusion probability, like perceived similarity, decreases exponentially with increasing distance (Shepard, 1987).

In the same way that temporal contiguity effect describes how limitations of temporal distinctiveness explains transition and transposition gradients in memory for lists of items, Rerko et al. (2014) refer to an analogous effect in the spatial domain to explain similarly graded effects of distance, in that spatial confusions between items are more common at smaller distances (Emrich & Ferber, 2012; Bays, 2016; Sahan et al., 2019). The link between swap errors in VWM and transposition errors in serial recall has been proposed to reflect a more general mechanism in memory by which items are bound to context dimensions (Oberauer & Lin, 2017; Schneegan et al., in press). The present study aims to further extend the Popov et al. (2021) findings by systematically modelling the rate at which intrusion probability decreases with increasing distinctiveness in the spatial domain, as well as the temporal domain. In addition to associations formed between items due to the environment in which they are studied, we also extend this line of reasoning to qualities of the items themselves. In the case of word stimuli, two examples of such qualities are semantic associations between words, and perceptual similarity in terms of the orthography of the words when visually presented.

### Semantic and Orthographic Similarity

Many source memory paradigms, including that of the present study, use word stimuli in which semantic associations are particularly salient. The Deese-Roediger-McDermott (DRM; Deese, 1959; Roediger & McDermott, 1995) paradigm is an influential demonstration of how semantic association between words can result in false memory for non-presented words, known as critical lures. When asked to make source judgements for critical lures, participants make source attributions with high confidence, corresponding to the source of the semantically related words (Lampinen et al., 1999; Gallo et al., 2001; Gallo & Roediger, 2003; Roediger et al., 2004). Furthermore, when a list of semantically associated words is split between two sources such that one source presents the strongest associations and the other presents the weakest, critical lures are consistently attributed to the source with the stronger half (Hicks & Hancock, 2002; Hicks & Starns, 2006).

Returning to the principle of contiguity in free recall, past research has demonstrated that in addition to recalling items that were presented close together in time, participants also demonstrate a tendency to successively recall items that are semantically related to each other, suggesting more broadly that items in memory are organized not only by contextual features, but also features of the items themselves (Glanzer, 1969; Howard & Kahana, 2002; Morton & Polyn, 2015). Notably, the tendency for semantic factors to influence retrieval has been observed not only when lists are specifically constructed with an obvious semantic theme (Bousfield, 1953; Puff, 1966), but also when lists are randomly constructed with seemingly unrelated words (Schwartz & Humphreys, 1973; Tulving, 1962; Howard & Kahana, 2002). Given that semantic associations between words can affect retrieval for even unrelated lists, our modeling of systematic variations in intrusions from non targets in the current study must account for item features in addition to contextual features.

Another feature that affects memory for words is their orthographic and/or phonological similarity (Conrad, 1963; Wickelgren, 1965). The effect of these features on false memory specifically has also been studied using the DRM paradigm. Sommers and Lewis (1999) constructed lists of phonologically related words and found that rates of false recall were highest when words were close phonological neighbors and lower when words were phonologically dissimilar, suggesting that words that differ by a single grapheme or letter were most likely to be confused. In the current study, words are presented visually, and for brevity, we refer exclusively to the orthographic similarity between words. Additional research has simultaneously investigated recall for lists of semantically and orthographically similar words, and suggests that distinct mechanisms drive false memory for each kind of similarity and that in mixed lists, the processing of semantic information is simultaneously integrated with orthographic information in memory retrieval (Massaro et al., 1991; Watson et al., 2002; Nieznański et al., 2019; Chang & Brainerd, 2021; Coane et al., 2021).In our investigation of how item features contribute to intrusion errors in the continuous-outcome source memory paradigm, we compare models in which semantic, orthographic, or both factors combine additively or multiplicatively with contextual factors.

## The Present Study

In this study, our primary aim was to develop and compare systematic models of intrusions in the continuous-outcome source memory task, using both response error and RT data. Over two experiments, we compare models in which intrusion probabilities are sensitive to temporal similarity between item pairs, to both spatial and temporal similarity, as well as semantic and orthographic features of the word stimuli. To manage the number of competing models, we did not pursue a full factorial approach to modelling intrusions. Instead, we build upon the Popov et al. (2021) finding of a temporal effect and introduce additional components in sequence, firstly extending the contextual similarity between items to include space as well as time, and then also introducing features of the words in terms of their semantic and orthographic similarity to each other, evaluating the improvement in model fit with each addition.

- This seems to be repeating stuff said earlier. I’ve recommended saying it right at the beginning. I am a firm believer in telling the reader what the study is about and why it is important as early as possible – basically start with a big-picture sketch, then go back and fill in the details and review the critical literature later, so the reader knows where the argument is going.

A secondary aim was to investigate if the simultaneous presentation of information (i.e. the presentation of a word in the location) or the sequential presentation of source and item information (i.e presenting the source location, followed by the item as in Harlow & Donaldson, 2013; Zhou et al., 2021) affected source responding. To this end, we experimentally manipulated the format of source presentation across participants. Our broader research goal, under which both of these aims fall, is to assess whether previous characterizations of source memory retrieval as a thresholded process hold when changes in stimuli presentation and the possibility of intrusions from non target items are considered.

In Experiment 1, in which we collected data from a large number of participants who each completed three experimental sessions, we found qualitative improvements in model fit by introducing successively more elaborated models of intrusions between items, ranging a pure guessing model with no intrusions to a model in which the intrusion probability was determined by a spatiotemporal and word feature similarity gradient. However, the quantitative evidence for the spatiotemporal gradient model was inconclusive, which may have been due to an insufficient number of observations reflecting intrusion responses to support the parameter penalty incurred by the more complex models. In Experiment 2, we address this issue by concentrating power at the level of individuals by using a small-*N* design which found that a spatiotemporal intrusion model was quantitatively preferred, supporting the view that spatiotemporal similarity influences intrusion probability, but did not find support for contributions from semantic and orthographic similarity.

# Experiment 1

## Method

### Participants

Experiment 1 collected data from 50 participants, each of whom served in three experimental sessions, each of around one hour duration [?]. Ten were recruited online through the University of Melbourne undergraduate research experience program and 40 s were recruited via *Prolific*, an online participant recruitment platform, each of whom served in three experimental sessions. Five participants from the undergraduate pool and seven participants from the Prolific pool did not complete all sessions of the online experiment, resulting in incomplete datasets which were excluded from the final analyses. Additionally, two participants recruited via Prolific were excluded due to at-chance performance in the memory retrieval task, measured by applying the Rayleigh test which indicated no evidence for a departure from uniformity, interpretable as completely random responding. After exclusion, there were five undergraduate participants and 31 Prolific participants, for a total sample of 36 participants. For their participation in each session, undergraduate students were granted credit towards course requirements, and Prolific participants were paid 6.50 GBP/hour. Participants were provided with plain language statements and consent forms and gave informed consent prior to the start of the first session of the experiment.

### Stimuli and Apparatus

Stimuli were low-frequency, four-letter words from the SUBTLEXus database. (reference?), The frequencies ranged from 1 and 300, which represents the number of times the word appears in the corpus of 51 million words. Words were displayed in 24 point Courier New white font positioned in the center of a uniform gray [?] mean luminance field. The use of a monospaced font and the restriction to four letters ensured that stimuli always occupied the same amount of space on the screen. Software written in JavaScript using the jsPsych presentation language [?] (de Leeuw, 2015) controlled stimulus presentation and recorded responses. [Worth making the point, but a bit more compactly?] de Leeuw and Motz (2016) compared the accuracy of RTs recorded using JavaScript and under laboratory conditions using Psychophysics Toolbox and found that the JavaScript introduced a small and consistent measurement bias: RTs recorded under Javascript were around 25 longer than under Psychophysics Toolbox, but there were no systematic differences in RT variability. Biases of this magnitude are negligible for the purposes of the inferences we wish to draw about RTs in our task.

To address potential concerns about the precision of web-based RT measurements compared to typical laboratory conditions, a comparison of RT measurements collected using JavaScript and the more traditional Psychophysics Toolbox (PTB) by de Leeuw and Motz (2016) found that although the former measured RTs approximately 25 ms longer than the latter, this difference was uniform across measurements and there was no reliable difference in the variability of RT distributions. We observed that RTs in the current study were generally in accordance with those of Zhou et al. (2021) which used a similar source memory paradigm with PTB in a laboratory setting.

### Procedure

Participants completed the experimental tasks over three sessions. Each of the three sessions consisted of 120 trials, presented in 12 blocks of ten items each. Each block consisted of a study phase, a mathematics distractor phase, a recognition phase, and finally a source recall phase. There were additionally five practice trials at the beginning of each session, the data from which was not included for analysis. Presentation format was manipulated between participants, with participants randomly allocated to either a simultaneous study condition or a sequential study condition, which remained the same across experimental sessions for each participant. All other phases were identical between the conditions.

In the sequential study condition, participants were presented with a black marker positioned on a randomly generated angle on the outline of a circle at the start of each trial for 600 ms. The presentation of the marker was followed by the display of a word in the center of the screen for 1500 ms. To ensure that participants attended to the source information, they were instructed to indicate the previous location of the cross on the blank target circle using a computer mouse. Responses made within π/8 radians of the true target location were classified as attended and advanced participants to the next item. Responses further away were deemed unattended and the words “TOO DISTANT” was displayed for 1000 ms, then the location was then re-presented and the verification task was repeated.

In the simultaneous study condition, participants were presented with the marker and the word simultaneously for 1000 ms. Instead of being positioning the word in the center of the screen, in the simultaneous encoding condition, the word was positioned at the same angle as the marker, offset by a longer radius. The location of the word relative to the marker was determined by the sector the angle was in, with the word being offset to one of eight points on the bounds of the text box, corresponding to the middle of each of the four sides, and the four corners (i.e. in the North sector, the anchor was the bottom middle of the text box, while in the Northeast sector the anchor was the bottom left of the text box). As with the sequential condition, a verification task followed each presentation, which was repeated until participants reproduced the location to within π/8 radians of the presented angle.

After studying each of the items for that block, participants were then instructed to complete a distractor task, which involved 30 seconds of arithmetic problems. These problems were presented as three single digit integers, which summed to a fourth number which would either be the correct sum, or a number that was one higher or lower than the actual sum. Participants would indicate if the sum was correct by pressing the keys 0 (false) or 1 (true).

In the recognition phase, participants were shown a shuffled list of 10 previously studied items and 10 foils and asked to rate each item on a six-point Old/New confidence scale. Participants responded by pressing a number from 1 to 6 on their keyboard, with 1 representing “Sure New” and 6 representing “Sure Old”.

Finally, in the source memory retrieval task, participants were cued with the words for 1500 ms, and then indicated the recalled location by a moving the mouse from the starting point in the centre of the circle to a point on the circumference of the response circle. Response time was measured from the first movement of the mouse beyond a calibration marker, which was a circle with a radius of 8 pixels in the center of the screen. The cursor was required to be centered on this calibration marker to begin each trial. There was no time limit on the decision task. A schematic for one trial in each of the phases is shown in Figure 3.

Figure 3

*Schematic of one Trial in each Phase of the Experimental Paradigm.*

A picture containing clock

Description automatically generated

## Results

In this section, we compare several models of response errors and then repeat the analysis using the circular diffusion model in which we compare models of both response errors and RT. The purpose of the two sets of analyses was to ascertain whether the inferences we draw about intrusion processes are altered if the models are required to account for RT as well as accuracy. For both sets of models we investigate intrusion models of varying complexity. We compare pure guessing and pure intrusions models with a hybrid model that incorporates both intrusions and guessing, and then consider more sophisticated models in which the intrusion probabilities depend on the temporal, spatiotemporal, or a combination of spatiotemporal and semantic and/or orthographic similarity between items. Prior these analyses, we assess whether recognition and source judgments vary with sequential versus simultaneous item presentation.

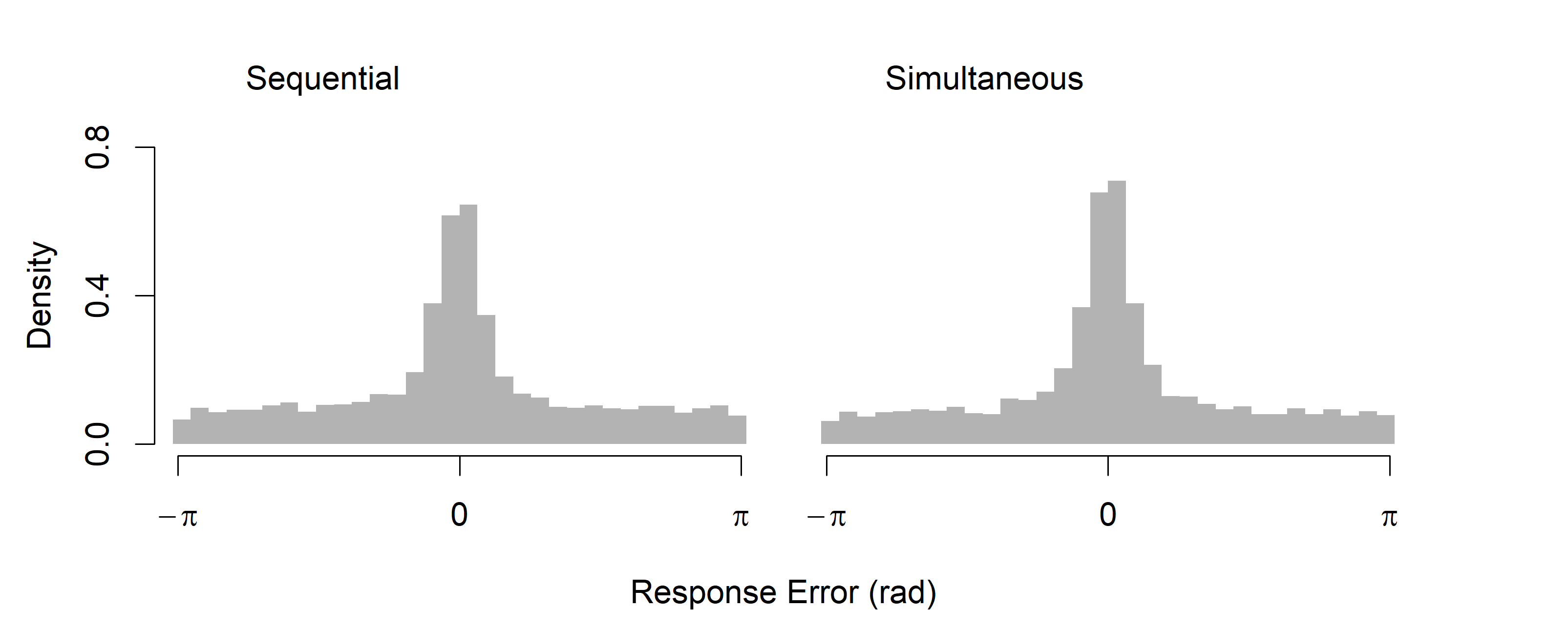
### Data Exclusion

In addition to the previously described exclusion of two participants’ data, individual responses from the remaining participants with a response time of faster than 300 ms or slower than 7000 ms were also excluded from subsequent analyses. This resulted in the omission of 1.72% of data.

### Simultaneous and Sequential Presentation

To assess whether presentation format influenced performance in the source retrieval task, we pooled data across participants in each condition. The distributions of response error in both conditions are characteristically leptokurtic in shape, with tall central peaks and heavy tails (Figure 4).

Figure 5   
*Normalized Histograms of Source Error in Sequential and Simultaneous Presentation Conditions Pooled across Subjects in each Condition*



Is this the correct way to do this analysis? Where does df = 12556 come from? It looks as though you’ve just pooled all of the trials for all of the participants and done a t-test on them, which is treating the data as 12,558 independent observations, which is not correct. The way I would do this is to calculate a mean absolute error for each participant and then do a t-test on that. My strong suspicion is that this test won’t be significant. The more complex way of doing the test would be as a repeated-measures ANOVA in GLM, which had conditions and repeated trials for each participant as factors, but it should give the same result because the main effect of condition is simply testing for differences in the means of the individual observations for each participant in the two conditions.

Although the distribution of errors in the two conditions are visually similar, Welch’s *t*-test indicated that the mean absolute error for participants in the simultaneous condition (M = .98, SD = .95) was significantly lower than mean absolute error in the sequential condition (M = 1.05, SD = .97), *t*(12556) = 4.31, *p* < .001. In our subsequent modeling analyses, we fit data from each participant separately, and for the most part we did not find significant differences between individual-level parameter estimates across conditions. These analyses are provided as supplementary material. Ultimately, although presentation format appears to influence source performance in terms of mean error, the general shape of the distributions of errors, in particular the presence of heavy tails, is the same in both conditions. For the purposes of our broader question of whether source memory retrieval is thresholded, it is clear that the heavy tails, which are often found to indicate a guessing process, are not a byproduct of the presentation format. For this reason, we do not make further reference to the presentation manipulation in our subsequent modeling.

### Response Error Models

If this is going to a US journal, “modeling” rather than “modelling.” Ugly, but we have to live with it!

Our modelling strategy was to start with a two-component mixture model similar to that of Zhang and Luck (2008), and then introduce successive elaborations of the intrusions model that represented different kinds of target-distractor similarity,, first temporal similarity, then spatiotemporal similarity, and finally semantic and orthographic similarity. The same stepwise process was followed with the circular diffusion model, using the same calculations to weight intrusion probability according to similarity, using the Zhou et al. (2021) two-component circular diffusion model as the decision model in instead of the Zhang and Luck (2008) error model. The models are formally described in the sections to follow. In addition, we implemented variations of some models that permitted different weightings for primacy and recency of intrusions and compared additive and multiplicative combinations of similarity. For ease of presentation, we have excluded these model variants in this text, but code for all the models is available at https://osf.io/76wtm/ and are provided as supplementary material.

### Model 1: Pure Guessing

As previously described, the Zhang and Luck (2008) model expresses the idea that responses are generated from a mixture of two process, the first process is target-driven responding that follows a von Mises distribution. The form of the von Mises probability density function with precision κ and mean μ is

|  |  |
| --- | --- |
|  | (2) |

where the normalizing constant is a modified Bessel function of the first kind of order zero, 8iuuwhich is expressed in the second equality to show that the exponential term in the numerator is normalized by the integral of the exponential term across the domain [0, 2π], the perimeter of the circle (Smith, 2016). Note that κ in this model is freely estimated, while in the circular diffusion models to follow, κ is determined by the quality of evidence and decision threshold as given in (1). Jason, it’s correct, but the integral is not elementary to evaluate (it uses residue calculus). If you want a reference, you can cite Equation 3.339 in Gradshteyn and Ryzhik (2007) (See Smith et al. (2020) for the complete reference.)

The distribution of target responses given by the von Mises distribution is mixed with a proportion of guesses which are distributed uniformly around the circle

|  |  |
| --- | --- |
|  | (3) |

where the probability that a response is a guess is represented by . In the target-driven component, represents the target angle, is the reported angle, and represents a von Mises distribution with a mean of 0 and with precision . The uniform component can alternatively be viewed as a von Mises distribution with zero precision*.*

### Model 2: Pure Intrusions

To test the strong prediction that all non target responses can be accounted for with intrusions from non target items without invoking guessing, Model 2 substitutes the guessing component in the mixture model with an intrusion component:

|  |  |
| --- | --- |
|  | (4) |

where the probability of an intrusion occurring is represented by , and the angle associated with the *i*th intruding item is represented by . Note that of the *m* non target items, the probability of a particular non target intruding is equal.

### Model 3: Intrusions + Guessing (Flat Gradient)

Model 3 combines intrusion and guess responses in the three-component model of Bays et al. (2009):

|  |  |
| --- | --- |
|  | (5) |

In contrast to subsequent models where the probability of a given non target item intruding is dependent on its similarity to the target, intrusions in Model 3 all occur with equal probability. We refer to this feature of the model as a flat intrusion gradient.

### Model 4: Temporal Similarity Gradient

In contrast to Models 2 and 3 in which each intrusion is equally weighted (that is, the likelihood of each intruding item is simply divided by the number of possible intrusions), in Model 4 the probability of each non target item intruding is determined by its temporal similarity to the target represented by *t:*

|  |  |
| --- | --- |
|  | (6) |

Is it worth including a figure showing the temporal similarity gradient model, which would emphasize the forward and backward asymmetry? It’s such an important idea in the paper it would be good to make it completely clear to readers the kind of function it is describing. It would also allow you show what numerical variation in the parameters does to the predictions.

We incorporate the assumption that the strength of association between items is an exponentially decreasing function of distance, represented by *l*, the lag of the intruding item from the target (Shepard, 1987). To allow for asymmetry in terms of temporal similarity for backwards and forwards lags, scales the similarity slope in each direction such that when , items presented after the target have greater temporal similarity, and hence are weighted more in calculating the overall likelihood of intrusion, compared to items preceding the target. The rate of exponential decay, , is estimated separately for the forwards and backwards similarity slopes.

The probability of an intrusion occurring on a trial is the sum of temporal similarity values over all the possible nontarget lags for the study list position of the target.

|  |  |
| --- | --- |
|  | (7) |

Temporal similarity, *t*, is subscripted to reflect the fact that each item in the study list has a unique similarity value which varies depending on its proximity to the target item. Because the possible lags are different for each position in the study list, the summed probability of intrusions also varies across trials. We assume that these changes in intrusion probability are reflected only in the probability of a target response, and not the probability of guessing *β* which is constant across study positions We also implemented alternative models where 1) the probability of memory responses was constant (and guesses were sensitive to summed intrusion probability), and 2) both guesses and memory changed across trials depending on an additional arbitrary mixture parameter, neither of which improved the fit of the model. We consider the plausibility of these assumptions and the limitation of mixture modelling this ambiguity reflects in the discussion section.

### Model 5: Spatiotemporal Similarity Gradient

Using the same basic structure as the previous models, in Model 5 intrusion likelihood is a weighted product of temporal and spatial (or locational) similarity:

|  |  |
| --- | --- |
|  | (8) |
|  | (9) |

where the overall weight given to each intruding angle, *w*, is determined by both the temporal similarity between the intruding item and the target, *t* as defined in (6), and the spatial/locational similarity between the target and intruding angles *l*:

|  |  |
| --- | --- |
|  | (10) |

as with temporal similarity, we assume that spatial similarity decreases exponentially with distance, which in this case is the circular distance between the two angles. The relative contribution of temporal and spatial similarity in determining the probability of a particular non target item intruding is weighted by . When = 0, *w* = *t* in Equation 9 and the weight function is independent of spatial location, while when = 1, *w* = *l* and the weight function is independent of lag. At intermediate values of , *w* is jointly a function of the two. [I think it is worth spelling out the implications of Equation 9 explicitly.] Naturally, intrusion responses from near nontargets will be associated with lower error relative to the target than intrusions from far nontargets. Therefore, as increases, overall response error decreases. In addition to the effects of spatiotemporal similarity on intrusion probability, we also introduce models that incorporate orthographic and semantic similarity as further elaborations to the model.

### Model 6: Orthographic Model

In the orthographic model, orthographic similarity between the target and a non target word is represented by *o* and is calculated from the Levenshtein distance of the two four-letter strings, and then weighted against the spatiotemporal similarity of the presentation context given in (9). Levenshtein distance measures the minimum number of edits (insertions, deletions, or transpositions) (correct? Need to say what an “edit” is. It’s not self-evident.) (because all strings were of equal length, all edits were substitutions) of single letters to transform one string into another. The resultant weight then determines the individual probability of a given non target item intruding

|  |  |
| --- | --- |
|  | (11) |

In words, the probability of a non target item intruding is a weighted product of the spatiotemporal similarity of the two items at presentation (which is itself a weighted product of temporal and spatial similarity), and the orthographic similarity between the non target and target word. I’m assuming Equation 11 is displaying wrongly in LibreOffice and its just Equation 9 generalized with an extra weighting term?

### Model 7: Semantic Model

The semantic model substitutes orthographic similarity in Model 6 for semantic similarity between target and non target words. To model semantic associations between words in our model, we used vector representations of each word, with each vector consisting of 300 internal dimensions, obtained from a *word2vec* model that was pre-trained on multiple corpora of natural text (Mikolov et al., 2017)[[4]](#footnote-5). Word2vec belongs to a class of models which predict relationships between words, and which have been found to outperform more traditional approaches that count co-occurrences between words in particular contexts such as Latent Semantic Analysis (LSA; Landauer & Dumais, 1997; Mander at al., 2016). Semantic similarity in our model, *s*, is defined as the cosine similarity between these vector representations, and is combined multiplicatively with temporal similarity, *t*, to give the weighted spatiotemporal similarity, *w*, for each non target item in each trial:

|  |  |
| --- | --- |
|  | (12) |

### Model 8: Four-Factor Model (Multiplicative)

In the four-factor model, both semantic and orthographic components are combined multiplicatively with an additional parameter, ψ*,* governing the weight of semantic similarity relative to orthographic similarity: As it appears to me here, Equation 13 doesn’t include ψ

|  |  |
| --- | --- |
|  | (13) |

We parameterized the four factors in a nested fashion to ease interpretation of each of the weights within the multiplicative combinations.

### Model 9: Four-Factor Model (Additive)

We also implemented models in which these different kinds of similarity combine additively. Model 9 expresses the idea that factors related to the (spatiotemporal) presentation of the words and factors of the word themselves are independent and are additive:

|  |  |
| --- | --- |
|  | (14) |

Other additive combinations were tried, both in the inner and outer combinations of the nested structure of the weights, but these models were generally not preferred over their multiplicative versions. All models under consideration and the corresponding parameters are summarized in Table 5.

Table 1

*Summary of Response Error Models and Parameters*

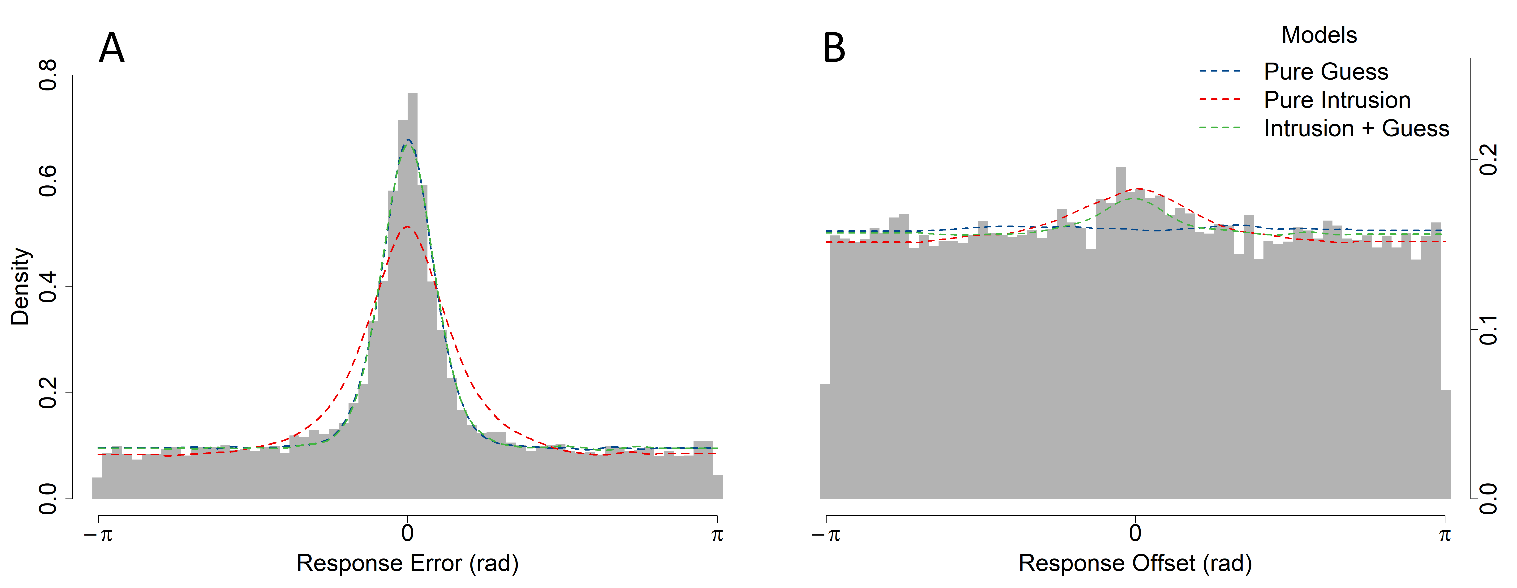
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Parameter** | | **Description** | | |
| κ1 | | Precision, memory | | |
| κ2 | | Precision, intrusion | | |
| β | | Proportion of uniform guesses | | |
| γ | | Proportion of intrusion responses | | |
| τ | | Temporal gradient asymmetry | | |
| λ1 | | Temporal similarity decay, forwards | | |
| λ2 | | Temporal similarity decay, backwards | | |
| ζ | | Spatial similarity decay | | |
| ρ | | Spatial vs. Temporal similarity weight | | |
| χ | | Spatiotemporal vs. Semantic/Orthographic weight | | |
| ψ | | Semantic vs. Orthographic weight | | |
| **Model** | **Parameters** | | **Number of Parameters** | |
| 1. Pure Guess | κ1, β | | | 2 |
| 2. Pure Intrusion | κ1, κ2, γ | | | 3 |
| 3. Intrusion + Guess (Flat) | κ1, κ2, β,γ | | | 4 |
| 4. Temporal Gradient | κ1, κ2, β,γ, τ, λ1, λ2 | | | 7 |
| 5. Spatiotemporal Gradient | κ1, κ2, β,γ, τ, λ1, λ2, ζ, ρ | | | 9 |
| 6. Orthographic Gradient | κ1, κ2, β,γ, τ, λ1, λ2, ζ, ρ, χ | | | 10 |
| 7. Semantic Gradient | κ1, κ2, β,γ, τ, λ1, λ2, ζ, ρ, χ | | | 10 |
| 8. Four Factor (Additive) | κ1, κ2, β,γ, τ, λ1, λ2, ζ, ρ, χ, ψ | | | 11 |
| 9. Four Factor (Multiplicative) | κ1, κ2, β,γ, τ, λ1, λ2, ζ, ρ, χ, ψ | | | 11 |

### Response Error Model Comparison Do a global search and replace of “non target” with “nontarget”

First, we compare Models 1, 2, and 3 to focus on how including a basic intrusion component where all nontargets are equally likely to intrude affects the predictions of the model. As established by Bays et al. (2009), although guesses and intrusions will both appear uniform relative to the target on each trial, the two can be distinguished by examining the distance between responses and each of the non target items on each trial. With no contribution of intrusions, the resultant distribution should appear uniform, while evidence for intrusions is reflected in the kind of central tendency present in our data as shown in Figure 5B. We will subsequently refer to this analysis as *recentering* the data, as it is equivalent to recentering the distribution of response errors on the non target angles.

Figure 6

*Comparison of two-component and Intrusion + Guess models' Predictions of Response Error.*



*Note.* Panel A shows the distribution of response errors, defined as the angular distance between the response angle for each trial with the target angle on that trial. In panel B, distances are instead calculated between the response angle and each non target angle, that is, the location of all other items in the block excluding the trial target. Observed data are represented by grey \_ If American journal will be “gray” histograms. histograms, while model predictions are represented by dashed lines.

It probably doesn’t matter for a first submission, but you will need to think about how your figures will reproduce in greyscale in a printed journal if you don’t have enough to pay for coloured figures (about $1,000 a figure!) I find it best to think about these things from the outset. Even in color, these figures are not particularly legible. They would be improved by changing the line weight and line stype. Also, is it the case the Pure Guess and Intrusion + Guess functions lie on top of each other? I would say that in the caption; otherwise people will go looking for the missing function.

While all three of these models capture the heavy tailed distribution of errors to a similar extent, the Pure Intrusion model underpredicts the precision of memory responses centered on 0, as seen in Figure 3A. In contrast, the Pure Guess model does capture both the tails and peak of the distribution of response errors, but because it does not predict any relationship between the non target and response angles, it fails predict the central tendency evident in the recentered data shown in Figure 3B. Of these models, only the Intrusions + Guess model, with both guessing and intrusion components, is able to produce both patterns of data at the same time, suggesting both processes are important to explain the distributions of response error. This point is further illustrated by the parameter values that resulted in the best fit of each model (Table 2 displays these values for all models averaged across all participants). Specifically, the addition of intrusions in the Intrusion + Guess model reduces the estimated rate of guessing relative to the Pure Guess model, but it does not eliminate guessing entirely. Notably, the Pure Guess and Intrusion + Guess models agree on the proportion of non target responses (β = 0.60 in the former, β + γ *≈* 0.60 in the latter). In Pure Intrusion model, with no guesses, the estimated precision of memory responses is lower relative to the alternative models, which is reflected visually in Figure 3A with the misprediction of wider shoulders and shorter peak compared to the data, which is well captured by the other models. This last para reads as a bit stream-of-consciousness. Can you tighten it up a bit?

Table 2

*Average Parameter Estimates for Each Model to Experiment 1 Data.*

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Parameter Average | | | | | | | | | | |
| κ1 | κ2 | **β** | γ | τ | λ1 | λ2 | ζ | ρ | χ | ψ |
| 1 | 19.53 |  | **0.60** |  |  |  |  |  |  |  |  |
| 2 | 5.31 | 4.28 |  | 0.46 |  |  |  |  |  |  |  |
| 3 | 19.06 | 14.64 | **0.36** | 0.24 |  |  |  |  |  |  |  |
| 4 | 16.02 | 10.10 | **0.39** | 0.28 | 0.56 | 0.89 | 1.08 |  |  |  |  |
| 5 | 18.82 | 8.86 | **0.39** | 0.22 | 0.56 | 1.69 | 1.55 | 0.52 | 0.63 |  |  |
| 6 | 14.66 | 12.67 | **0.49** | 0.42 | 0.59 | 1.97 | 2.18 | 0.53 | 0.40 | 0.34 |  |
| 7 | 16.84 | 11.27 | **0.54** | 0.19 | 0.59 | 1.64 | 1.50 | 0.39 | 0.46 | 0.50 |  |
| 8 | 16.26 | 11.48 | **0.59** | 0.31 | 0.58 | 1.21 | 1.48 | 0.38 | 0.43 | 0.15 | 0.22 |
| 9 | 14.21 | 12.33 | **0.51** | 0.46 | 0.45 | 1.82 | 2.02 | 0.51 | 0.29 | 0.34 | 0.24 |

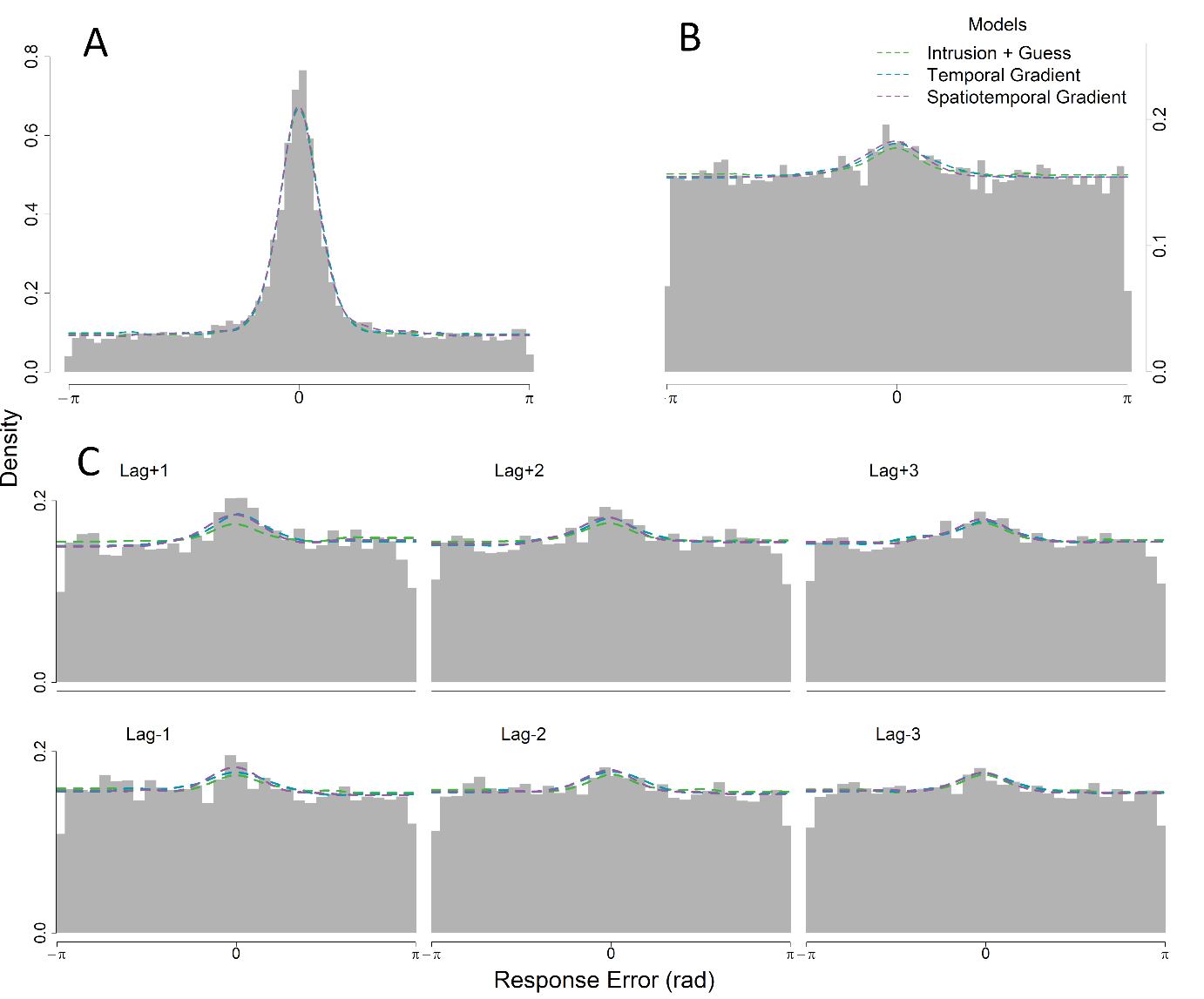
*Note*. The estimated proportion of guesses (β, boldface) decreases when comparing Model 1, with no intrusions, to Model 3 with flat intrusions. However, subsequent gradient elaborations on the intrusion component do not further decrease the estimated proportion of guesses.

Having established that intrusions do contribute to errors, we now seek to characterize the nature of the intrusion process more precisely. As discussed above, we did so by comparing models in which spatial, temporal, semantic, and orthographic similarity between targets and nontargets affected the intusion probability.

we now consider refinements of the intrusion component of the model so that the probability of intrusions systematically vary with similarity with the target item. The base Intrusion + Guess, Temporal Gradient, and Spatiotemporal Gradient models make visually indistinguishable predictions about the overall distribution of response errors (Figure 6A) as well as the overall recentered errors (Figure 6B). Instead, the effect of different intrusion probability gradients can be seen in the recentered data in Figure 6C, which splits the recentered data by the lag and direction of the intrusion for each trial. Central tendency, and hence evidence for intrusions, is stronger in the forwards direction and decreases with higher absolute lag, where lag is defined as the number of positions in the study list separating the two items. Because the Intrusion + Guess assumes that intrusions are equally likely from all non target items, there is no relationship between lag magnitude or direction and how pronounced the central tendency is in the recentered data. In contrast, the Temporal and Spatiotemporal models predict fewer intrusions from greater lags and from backwards lags, a pattern which is present in the data (Figure 6C). We also fit further elaborations to the intrusion component of the Spatiotemporal model (i.e. the Orthographic, Semantic, and Four-Factor models) which did not improve the fit of the model enough to outweigh the AIC penalty associated with the additional parameters of these models. We have excluded the predictions of these models from Figures 6 and 7 to more clearly illustrate the differences between the Intrusion + Guess, Temporal, and Spatiotemporal models.

Figure 7

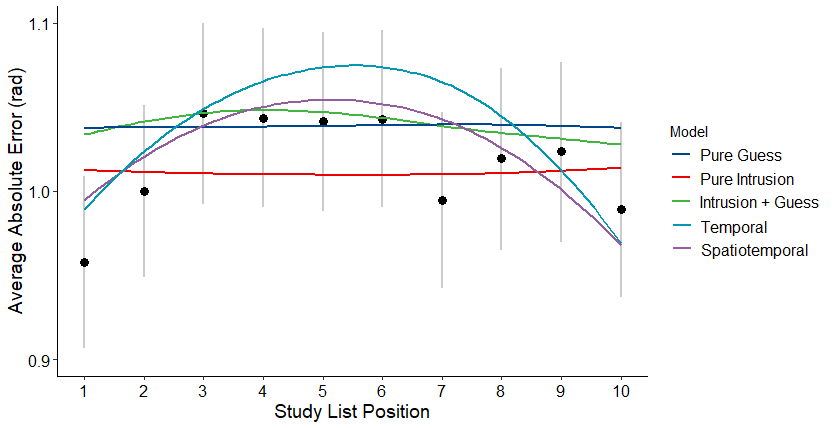
*Model Fits to Distances between Response Angles and Non target Angles by Direction and Lag*

*Note.* Panel A shows the distribution of response errors measured between the response angle and the target angle on each trial. Panel B shows the distribution of errors measured from the response angle and each of the non target angles in the studied list. Like Panel B, Panel C also shows errors relative to non target angles, but these distributions are conditioned on the lag, or the number of positions separating each non target angle and the target angle.

Another qualitative advantage of the Temporal and Spatiotemporal models over the Intrusion + Guess model is that they naturally predict a serial position effect, with lower response error for items at the start and end of the study list (Figure 5). Figure 8? The reason the gradient models make this prediction is because items at the beginning and end of the list have fewer close (small lag) neighbors, so they are less susceptible to intrusions from nontargets. the effect of the boundary at the ends of the list. For example, given that the greatest proportion of intrusions come from a lag of +1, then naturally the summed probability of intrusions is lowest for trials in which no items appear immediately after the target, i.e. the final trial in position 10. Furthermore, the temporal model overpredicts the strength of the serial position effect, particularly in overestimating errors for midlist items. In the spatiotemporal model, because intrusions from spatially closer non targets result in less response error, as measured from the target, overall response error is lower for the spatiotemporal model than for the temporal model, which in turn provides a better prediction of the pattern of average errors across study list positions.

Figure 8

*Average Response Error Across Target Serial Positions*



*Note*. Model predictions are represented by a loess curve through the average error of simulated data conditioned on serial position. Grey lines represent 95% confidence intervals, which were calculated using bootstrap sampling.

Despite these qualitative advantages of the intrusion gradient models over the flat Intrusions + Guess model, the latter is preferred in a quantitative sense on the basis of the Akaike information criterion (AIC). Models were fit on an individual level data, and the relative performance of the models summed over participants is shown in Table 3. Alongside the raw AIC values summed over participants, we also show the difference between each model and the best fitting model ΔΣAIC. T(need to define this notation?) hese values are also transformed into Akaike weights, *w*(AIC), which are interpretable as conditional probabilities for each model (Wagenmakers & Farrell, 2004). For the majority of participants, the Intrusion + Guess model is heavily preferred over the two-component Pure Guess and Pure Intrusion, as well as over the elaborated models with the various intrusion gradients. Individual *w*(AIC) comparisons are provided as supplementary material. Need a general qualitative summary here? “These results show that intrusions contribute significantly to the distributions of errors in our source retrieval task, but that, even when intrusions taken into account, a significant number of responses are best characterized as guesses, as predicted by a thresholded retrieval model.”

Table 3

*AIC Values Summed Over Participants*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Model Name | Parameters | ΣAIC | ΔΣAIC | | *w*(AIC) |
| 1. Pure Guess | 2 | 37338.77 | 276.86 | | 0 |
| 1. Pure Intrusion | 3 | 38178.07 | 1116.16 | | 0 |
| 1. Intrusion + Guess (Flat) | **4** | **37061.91** | **0** | | **1** |
| 1. Temporal Gradient | 7 | 37176.82 | 114.91 | | 0 |
| 1. Spatiotemporal Gradient | 9 | 37237.68 | 175.77 | | 0 |
| 1. Orthographic | 10 | 37633.28 | 569.28 | 0 | |
| 1. Semantic | 10 | 37310.81 | 246.81 | 0 | |
| 1. Four Factor (Additive) | 11 | 37363.50 | 299.50 | 0 | |
| 1. Four Factor (Multiplicative) | 11 | 37705.13 | 641.13 | 0 | |

This discrepancy between the quantitative fits and the qualitative patterns predicted by the models makes drawing a clear conclusion on the basis of the response error data difficult. To assess the models in a richer data space, we implemented diffusion analogs of each of the models, which we fit to both error and RT data.

In general yes, maybe needs elaboration? “The serial position dependence of response errors in Figure 8 favors a temporal gradient intrusion model, but the best model in an AIC sense Table 3 was the flat intrusion model because it required appreciably fewer parameters. To try to reconcile this discrepancy between the qualitative and quantitative fits we refit the joint error and RT distribution data using the circular diffusion model. The joint distribution data are considerably richer than the error data alone and impose much greater constraints on the models. In the recognition memory literature, models that have been difficult to distinguish using only accuracy (ROC) data have been successfully distinguished using sequential-sampling decision models that predict both RTs and accuracy (refs). Our aim in fitting the data with the circular diffusion model was to ascertain whether the use of joint error and RT data would allow us to distinguish between different versions of the intrusions model that are hard to distinguish in error data alone.

### Circular Diffusion Models

The parameterization of the full intrusion diffusion model is as follows: mean drift is represented by μ, which is normally distributed with standard deviation η*,* which reflects across-trial variability in evidence quality. We assume that memory strength differs between target and non target responses, and so these parameters were estimated separately for the memory component (μ1, η1) and the intrusion component (μ2, η2), however, the two components share a single decision criterion (*a*1) because we make the selective influence assumption that decision criteria should be unaffected by the identity of the stimulus.

A property of the circular diffusion model, inherited from the diffusion model of two-choice decisions (Ratcliff, 1978; Ratcliff & McKoon, 2008) is that it predicts slow errors. In difficult two-choice decisions in which accuracy is stressed, error RTs are typically slower than correct RTs (Luce, 1986, p. 233). The circular diffusion model makes a continuous counterpart of the slow-error prediction when drift rate varies across trials: The fastest responses are those made with the smallest error and RTs systematically increase with increasing error. In fits of the model to data, variability in drift rate norm is the most important source of variability needed to capture the distributions of error found in perceptual tasks (Smith et al., 2020) and memory tasks (Zhou et al. 2021). In the intrusion models, not only do drift rates vary across trials, but intrusion responses are assumed to be associated with lower mean drift rates than target responses, and so the prediction of a slow error effect is also a consequence of how intrusions are represented in the models. The uniform guessing component was implemented as a third diffusion process with a mean drift of 0 and a separate decision criterion (*a*2), reflecting a state in which no information is driving the decision process, which requires less total evidence to generate a response than information-driven trials. Finally, non-decision time (*Ter*) is added to response times to represent the assumption that RTs are the sum of the duration of the decision process as well as other processes, such as encoding and the response itself. For a more detailed description of the circular diffusion model, see Smith (2016). The parameters governing the mixture of memory, guess, and intrusion components are the same as in the response error models previously described. The parameterization of the diffusion models, as well as the AIC values summed over all participants, are summarized in Table 3. To focus on a smaller set of candidate models, we have excluded the diffusion models with orthographic and semantic similarity gradients in this analysis, but we reintroduce them in Experiment 2.

Table 3

*Diffusion Model Parameterization*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Parameters (Number)** | **ΣAIC** | **ΔΣAIC** | **w(AIC)** |
| 1. Pure Guess | μ1, η1, *a*1, *a*2, *Ter*,β**(6)** | 47611.67 | 1819.00 | 0 |
| 2. Pure Intrusion | μ1, η1, μ2, η2, *a*1, *Ter*,γ**(7)** | 46512.06 | 719.39 | 0 |
| 3. Intrusion + Guess (Flat) | μ1, η1, μ2, η2, *a*1, *a*2, *Ter*,β, γ**(9)** | 45850.07 | 57.41 | 0 |
| 4. Temporal Gradient | μ1, η1, μ2, η2, *a*1, *a*2, *Ter*,β, γ, τ, λ1, λ*2* **(12)** | 45988.75 | 196.09 | 0 |
| 5. Spatiotemporal Gradient | μ1, η1, μ2, η2, *a*1, *a*2, *Ter*,β, γ, τ, λ1, λ2, ζ, ρ **(14)** | 45792.67 | 0.00 | 1 |

In contrast to the response error model comparison, which showed a preference for the Flat Intrusion + Guess model, the spatiotemporal diffusion model is preferred over the other diffusion model variants. [This is all Figure 9 now?] Figure 8A shows the graphical fits of the diffusion models to the response error data. Compared to the equivalent plot for the models fit to response error data alone in Figure XA, the diffusion models appear to capture the distribution of response error more poorly. This is because the parameters of the diffusion model need to account for the entire joint distributions of RT and error, which is a 2D rather than a 1D distribution. In addition to the fits of the model predictions to the distribution of response error and RTs in Figure 8A and 8B respectively, Figure 8C shows the joint distributions of errors and RT in the form of a bivariate quantile plot, which depicts how response time (depicted on the y-axis) varies with response accuracy (depicted on the x-axis). In Figure 8C, the observed data are represented by points, with position along the x-axis representing the error quantiles (in sequence the 0.1, 0.3, 0.5, and 0.9 quantiles) such that the leftmost points closest to the origin is the value under which the most accurate 10% of responses lie, and so on for the points moving rightwards along the x-axis. The vertical stacks of points represent the response time quantiles (0.1, 0.5, 0.9) for data conditioned on the corresponding level of accuracy: the leftmost stack collectively represents response times for the most accurate 10% of responses, and the bottommost point in that stack is the fastest 10% of these most accurate responses.

Figure 9

*Diffusion Model Fits to Response Error and Latency*

Chart

Description automatically generated *Note*. Grey lines represent the 95% confidence interval around the observed response time quantiles. The error quantiles, which are unlabeled in the figure, are the .1, .3, .5, and .9 quantiles moving from left to right along the x-axis. The error quantiles are the upper bound defining the bin of responses for the corresponding response time quantiles, which are stacked vertically, while the lower bound for each bin is the next-lowest error quantile.

The average estimated values of each parameter are shown in Table 4. As with the response error models, including intrusions in Model 3 reduces but does not eliminate guesses compared to the Model 1.

Table 4

*Diffusion Model Parameter Estimates*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | μ1 | μ2 | η1 | η2 | *a*1 | *a*2 | γ | β | τ | λ1 | λ2 | ζ | ρ | *Ter* |
| 1 | 1.78 |  | 0.21 |  | 2.78 | 1.68 |  | 0.62 |  |  |  |  |  | 0.18 |
| 2 | 2.41 | 2.14 | 0.26 | 0.58 | 1.86 |  | 0.44 |  |  |  |  |  |  | 0.16 |
| 3 | 3.32 | 2.70 | 0.30 | 0.32 | 2.10 | 1.29 | 0.27 | 0.28 |  |  |  |  |  | 0.19 |
| 4 | 3.25 | 1.78 | 0.24 | 0.27 | 2.07 | 1.21 | 0.31 | 0.38 | 0.49 | 0.80 | 0.94 |  |  | 0.19 |
| 5 | 3.51 | 2.32 | 0.19 | 0.29 | 2.17 | 1.26 | 0.16 | 0.35 | 0.64 | 0.87 | 0.76 | 0.39 | 0.56 | 0.19 |

## Discussion

In Experiment 1, there were three key findings we wish to highlight. Firstly, our comparison of sequential and simultaneous presentation of word and location indicated that although overall response error was lower when words were presented simultaneously, similarly heavy tailed error distributions were present in both conditions, which suggests that while the sequential format used by Harlow and Donaldson (2013) and Zhou et al. (2021) may have marginally increased the difficulty of the source retrieval task relative to a simultaneous format, this difference does not challenge prior conclusions that source memory retrieval is a thresholded process. [But the failure to explicitly model intrusions does challenge them, doesn’t it? This is the whole point of your intrusion analysis. Once again, the intrusions story is the more important one theoretically, so I’d lead with that. And I’d phrase the methodological point somewhat differently. Something like: “...the difficulty of the continuous-outcome task and the distributions of errors it produces are not an artifact of the requirement to form an association between the item and its source location at encoding. Even when that association is made explicitly, by presenting the item in the source location, the distribution of errors is (virtually?!) unaltered.” [Maybe won’t need the qualification after redoing the t-test?] –

Secondly, we have found clear evidence that intrusions from non targets contribute to errors in our source memory task. The inclusion of intrusions in both the response error and diffusion models reduced the estimated proportion of guesses relative to the Pure Guess model. Our finding in the present study suggests that previous threshold models similarly overestimated guessing rates (Zhou et al., 2021). However, the poor fit of the Pure Intrusion models, again both in terms of error and joint error-RT data, suggests that a purely continuous view of source memory retrieval is incompatible with the data, even when intrusions are accounted for.

Thirdly, we found mixed evidence for systematic changes in intrusions based on similarity with the target item. Contrary to our expectations, elaborations of the intrusion component to model these systematic changes did not further reduce the estimated proportion of guesses in our model. We found successive qualitative improvement when intrusion probabilities were determined by temporal and spatiotemporal similarity gradients, compared to the base three-component Intrusion + Guess model in which all non target items are equally likely to intrude. However, when fit to response error data alone, the overall likelihood of the three-component models were sufficiently similar that the marginal improvements obtained with the more elaborated gradient models were outweighed by the parameter penalty associated with the gradients. When diffusion analogs of each model were additionally constrained by also fitting RT data, the model predictions were differentiated, resulting in a quantitative advantage for the spatiotemporal gradient. One explanation for the mixed results, both in terms of qualitative and quantitative response error evidence, as well as response error and joint error-RT data, is that there were simply insufficient observations in the participant-level data to support tests of complex models of intrusion effects on the basis of response error data alone. This motivated our use of a small-*N* design to concentrate power at the participant level in Experiment 2. When large numbers of trials are collected for individual participants each participant essentially becomes an independent replication of the experiment (Smith & Little, 2018). These kinds of small-*N* design have proven to be powerful tools for testing between complex models of decision making and other cognitive processes (Smith et al., 2020).

# Experiment 2

## Method

The experimental procedure for Experiment 2 was identical to Experiment 1 with the exceptions detailed below.

### Participants

Ten participants were recruited via Prolific, each of whom was initially recruited to serve in 10 experimental sessions. Four of the participants did not finish all 10 sessions and one was excluded because the Rayleigh test indicated no deviance from uniform responding, leaving a final sample of five participants included for the analyses.

### Procedure

Source and item information was only presented simultaneously for all participants in Experiment 2. The number of sessions each participant completed was increased from three sessions to 10 sessions. [This is unnecessary, as it’s made explicit in the previous section.] To maximize the number of trials at each serial position, in order to maximize the diagnosticity of our model selection procedure, we presented source and item information simultaneously on all trials of the experiment.

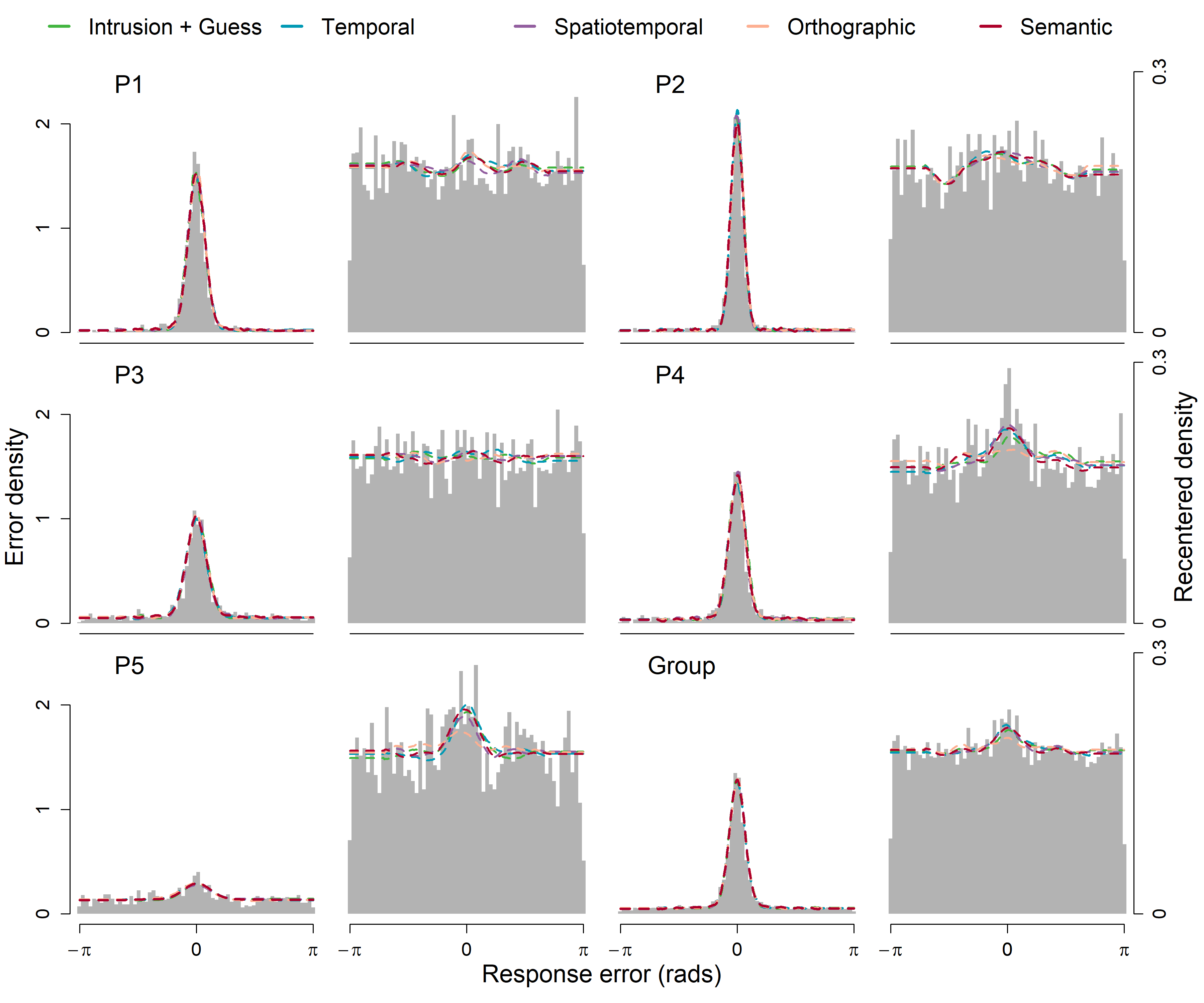
## Results

Experiment 1 clearly showed that a model with both intrusions and guessing was needed to produce the patterns of marginal and recentered response errors. We therefore excluded the Pure Intrusions and Pure Guess models and focused instead on models with both guessing and intrusions components.

In modeling the data, we excluded the Pure Intrusion and Pure Guess models because results from Experiment 1 clearly showed that a model with both intrusions and guessing is needed to produce the patterns of response errors and recentered errors in our task. In this section, we compare in more detail models in which the intrusion probabilities also depended on the orthographic and semantic similarity between targets and non targets. Figure 6 shows the graphical fits of the models to each participant-level dataset.

Figure 10

*Individual and Group-Level Fits of Models to Response Error and Recentered Error*



As with the response error models in Experiment 1, the error predictions of the models in Figure 6 are difficult to distinguish. One concern in comparing models is the diagnosticity of model selection when the models make similar predictions. To evaluate the extent to which our models mimic each other, we conducted a model recovery exercise, which we limited to the spatiotemporal, orthographic, and semantic models. We restricted this exercise to the most complex models as these were the ones most likely to lead to parameter tradeoffs and therefore be difficult to identify. The parameter values for each model that resulted in the best fit to each participants’ data was used to generate five simulated datasets for each participant, each with the same number of observations as the empirical dataset for that participant. Each simulated dataset was then cross-fit with the same set of models, and using the AIC as the fit statistic, we observed the number of times that the generating model was recovered as the best fitting model. Across all the simulated datasets, the spatiotemporal and orthographic models were successfully recovered in 80% and 84% of cases respectively. However, the semantic model was not recovered in any of the simulated datasets, for which the spatiotemporal model was universally preferred. The likely reason for this failure was because the semantic similarity of the stimuli was not systematically varied when we generated them and the average similarity between items was consequently low. Because the effect of semantic similarity is minimal in this dataset, the estimated value of χ is so low that simulated data generated from the fitted parameters are not distinguishable from the spatiotemporal model (average parameter estimate values are presented in Table 7).

Table 5

*Parameter estimates for each model, averaged across participants*

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Parameter Average | | | | | | | | | | |
| κ1 | κ2 | β | γ | τ | λ1 | λ2 | ζ | ρ | χ | ψ |
| 3 | 22.14 | 12.49 | 0.21 | 0.14 |  |  |  |  |  |  |  |
| 4 | 20.83 | 11.12 | 0.22 | 0.35 | 0.59 | 0.66 | 0.36 |  |  |  |  |
| 5 | 22.93 | 10.15 | 0.20 | 0.08 | 0.79 | 2.03 | 1.11 | 0.58 | 0.80 |  |  |
| 6 | 23.31 | 11.05 | 0.19 | 0.16 | 0.74 | 2.07 | 0.49 | 0.81 | 0.19 | 0.23 |  |
| 7 | 23.40 | 10.97 | 0.20 | 0.12 | 0.66 | 1.55 | 0.80 | 0.81 | 0.28 | 0.10 |  |
| 8 | 22.19 | 10.37 | 0.20 | 0.04 | 0.75 | 2.30 | 2.34 | 0.83 | 0.32 | 0.37 | 0.20 |
| 9 | 21.93 | 15.88 | 0.30 | 0.20 | 0.75 | 1.89 | 1.55 | 0.88 | 0.34 | 0.50 | 0.01 |

At an individual level, the spatiotemporal model is quantitatively preferred for majority of participants to varying degrees (Table 6). The balance of evidence in favor of the spatiotemporal model was strongest for Participant 1. While the spatiotemporal model is also preferred for Participant 4, the temporal and orthographic model are more competitive. The orthographic model is preferred outright for Participant 2, while for Participant 5 the simplest three-component model with uniform intrusion probabilities is preferred. The models in which intrusion probabilities are affected by semantic similarity (7, 8, and 9) were not well supported for any of the datasets.

Table 6

*AIC Weights for Individual and Group-level Response Error Model Fits*

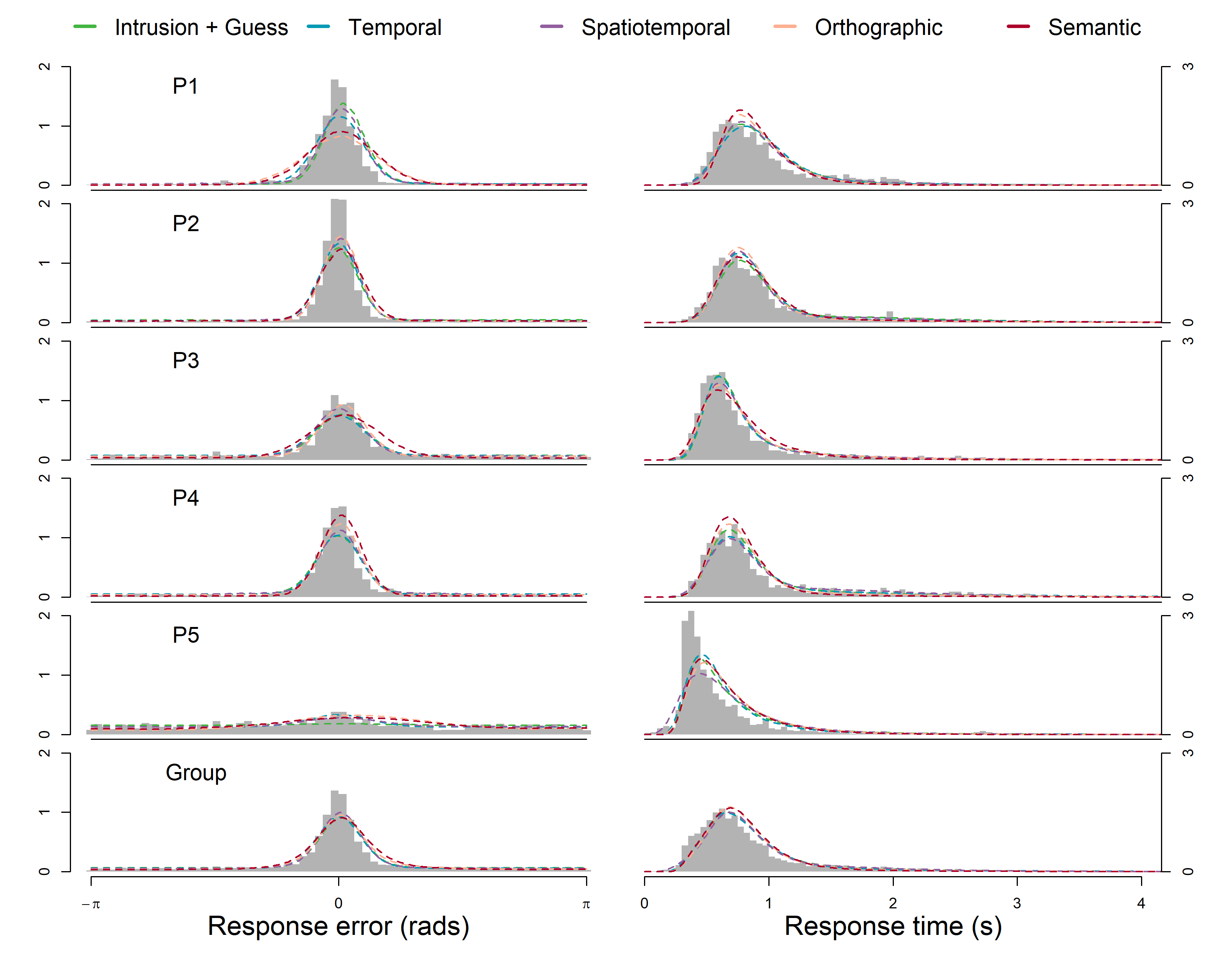
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Participant | Model | | | | | | |
| 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|  | AIC | | | | | | |
| 1 | 1116 | 1112 | **1088** | 1095 | 1096 | 1101 | 1097 |
| 2 | 555 | 558 | 562 | **550** | 563 | 556 | 559 |
| 3 | 2363 | 2364 | **2361** | 2367 | 2368 | 2370 | 2371 |
| 4 | 1520 | 1499 | **1498** | 1498 | 1501 | 1508 | 1577 |
| 5 | **3557** | 3567 | 3566 | 3567 | 3567 | 3570 | 3612 |
|  | w(AIC) | | | | | | |
| 1 | 0 | 0 | **.94** | .03 | .02 | 0 | .01 |
| 2 | .08 | .02 | 0 | **.82** | 0 | .06 | .01 |
| 3 | .23 | .14 | **.57** | .03 | .02 | .01 | 0 |
| 4 | 0 | .25 | **.39** | .28 | .08 | 0 | 0 |
| 5 | **.96** | .01 | .02 | .01 | .08 | 0 | 0 |

### Diffusion Models

As with Experiment 1, we also compared circular diffusion versions of each model. The graphical fit of the models to response error and times is shown in Figure 10. The additive and multiplicative four-factor models (models 8 and 9) performed worse than the three-factor semantic and orthographic models and are excluded from Figure 10 to better present the other model predictions. You probably need to comment, both in relation to Experiment 1 and 2, why the marginal error distribution fits of the diffusion model look worse. You need to spell out that model is attempting to fit joint RT distribution and accuracy data, whereas the von Mises analysis fits only the marginal error data. The poorer marginal error fits of the circular diffusion model are a reflection of the fact that these distributions are the marginals of the joint distributions and the model is attempting to account for all of the structure in the joint distributions simultaneously.

Figure 11

*Diffusion Fits to Participant-Level Response Error and Response Time Distributions*



The joint relationship between response error and RT is most clearly demonstrated at a group level in Figure 15, which plots response time quantiles for the data binned by response error quantiles. The semantic and orthographic models underestimate the number of responses made with high error. The average estimate for the proportion of guesses β is lower for these models (Table 9), resulting in a 0.9 error quantile prediction lower than the observed data or the models without orthographic or semantic intrusion factors. Conversely, the flat and temporal models overestimate high error responses, attributable to high estimates of β. The spatiotemporal model makes the closest prediction in terms of high error responses. Notably, all models under consideration misfit the .9 RT quantiles for the three most accurate error bins. This can be interpreted as a proportion of accurate responses which are slower than the models predict. In particular, the flat, temporal, and spatiotemporal gradient models overpredict the magnitude of the slow error effect, while the orthographic and semantic models uniformly underpredict the slowest RTs across the entire range of errors. There are substantial changes in RT across sessions, such that RTs in the first sessions tend to be slower than later sessions for most participants, which may explain why we do not observe a miss of this magnitude in Experiment 1, which had fewer subsequent sessions. [I get lost in relating this discussion to the list of parameters in Table 5. What is the difference between Models 8 and 9? Are these additive and multiplicative versions of the same models? If so, is it worth presenting both or is it just confusing? If I understand models properly, if a component goes to zero in a multiplicative model it should reduce to a simpler model without that component, but if it is not well identified in the data then you might end up long ellipses in the likelihood surface where the parameters are not well identified.]

Figure 12

*Model Fits to Group-level Joint Response Error and Time Quantiles*

Chart, scatter chart

Description automatically generated

Table 7

*Average Parameter Estimates*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Average Parameter Estimates | | | | | | | |
| μ1 | μ2 | η1 | η2 | *a*1 | *a*2 | γ | β |
| 3 | 3.95 | 1.51 | 0.32 | 0.39 | 2.83 | 1.36 | 0.12 | 0.30 |
| 4 | 4.34 | 0.97 | 0.04 | 0.10 | 2.78 | 1.44 | 0.14 | 0.32 |
| 5 | 4.71 | 1.46 | 0.24 | 0.08 | 3.03 | 1.34 | 0.07 | 0.27 |
| 6 | 3.71 | 1.76 | 0.44 | 0.19 | 2.60 | 1.37 | 0.17 | 0.20 |
| 7 | 3.18 | 1.04 | 0.41 | 0.15 | 2.30 | 1.21 | 0.15 | 0.15 |
| 8 | 3.69 | 1.72 | 0.27 | 0.24 | 2.70 | 1.49 | 0.12 | 0.15 |
| 9 | 4.26 | 0.16 | 0.11 | 0.01 | 2.72 | 1.29 | 0.09 | 0.34 |
|  | τ | λ1 | λ2 | ζ | ρ | χ | ψ | *Ter* |
| 3 |  |  |  |  |  |  |  | 0.10 |
| 4 | 0.49 | 1.23 | 0.53 |  |  |  |  | 0.07 |
| 5 | 0.72 | 0.17 | 0.62 | 0.78 | 0.86 |  |  | 0.07 |
| 6 | 0.67 | 0.77 | 1.12 | 0.36 | 0.51 | 0.23 | 0.00 | 0.11 |
| 7 | 0.70 | 0.62 | 0.89 | 0.29 | 0.22 | 0.33 | 1.00 | 0.13 |
| 8 | 0.62 | 0.45 | 0.94 | 0.37 | 0.48 | 0.18 | 0.38 | 0.08 |
| 9 | 0.74 | 0.70 | 1.03 | 0.10 | 0.45 | 0.46 | 0.10 | 0.09 |

Table 8 shows the AIC and AIC weights for the individual-level diffusion fits. The fit statistics support our qualitative comparison of the models: the spatiotemporal diffusion model is preferred for all five participants, although the difference in quality of fit is smaller between the flat gradient model and the spatiotemporal gradient model for Participant 1 than for the other participants.

Table 8

*Experiment 2 Diffusion Model AIC Comparison*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Participant | Model | | | | | | |
| 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|  | AIC | | | | | | |
| 1 | 2023 | 2094 | **2022** | 2540 | 2450 | 2063 | 2173 |
| 2 | 2016 | 2023 | **1912** | 2008 | 2115 | 2014 | 2063 |
| 3 | 3289 | 3277 | **3216** | 3243 | 3415 | 3271 | 3243 |
| 4 | 2991 | 2952 | **2790** | 2950 | 2849 | 2873 | 3287 |
| 5 | 4170 | 4284 | **4104** | 4398 | 4257 | 4145 | 4324 |
|  | w(AIC) | | | | | | |
| 1 | .43 | 0 | **.57** | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | **1** | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | **1** | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | **1** | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | **1** | 0 | 0 | 0 | 0 |

## Discussion

In Experiment 2, we found an overall preference in AIC terms for the spatiotemporal model. When the models were fit to response error data, the data from three of five participants were best fit by the spatiotemporal model, with the exceptions of Participant 2 and Participant 5, who illustrate the range of individual differences when the data are considered at a participant level. [awkward sentence!] Participant 2 is notable for having low response error overall and fewer non target responses but potentially being more sensitive to the word-based similarity between items when intrusions do occur, as suggested by a preference for the orthographic model. In contrast, Participant 5 responded with greater overall error, and was better fit by the flat intrusion gradient model.

When comparing the diffusion variants of the models, all five participants are best fit by the spatiotemporal model. The main reason for the spatiotemporal model’s advantage over the alternative models was its ability to predict the RT and the proportion of guessed responses with high error compared to the temporal and flat gradient models which overpredicted high error responses. In addition to the lower estimate for the proportion of uniform guesses, intrusions in the spatiotemporal model are also more likely to come from near non targets, which reduces the contribution of further intrusions which yield high error responses.

Contrary to our expectation that the similarity-based intrusion component in our model would be improved by adding item-based similarity to the model in the semantic and orthographic components, we did not find an advantage when comparing these models to the spatiotemporal model. One explanation for our finding comes from our choice in stimuli: words were limited to be exactly four letters in length, which limited the number of close semantic and orthographic word pairs. Additionally, study lists were constructed by randomly selecting words from across the entire stimuli pool, making high pairwise similarity within a single list even less likely, further limiting the potential effect of item-based similarity relative to the similarity of the spatiotemporal presentation context. While Sommers and Lewis (1999) found greatest confusability between words separated by a single grapheme, there were very few occasions in our experiment where a Levenshtein distance of 1 (an equivalent orthographic measure) occurred. However, it is worth noting that even subtle effects in semantic similarity in free recall tasks have been observed to have large effects on transitional probabilities in the sequence of recalled items (Howard & Kahana, 2002). That we did not observe an effect of semantics may be due to the particular demands of the source task, in that the location of the item was the reported feature. In addition, there is evidence from the visual working memory literature that location occupies a privileged role in the memory for item features (Pertzov & Husain, 2014). It is possible that using stimuli where word similarity are explicitly controlled, we might find a preference for the more elaborated intrusion models, but it is clear that spatiotemporal similarity dominates the probability of intrusions in the present dataset.

# General Discussion

Our goal in this study was to evaluate whether previous characterizations of source memory retrieval as a thresholded process (Harlow & Donaldson, 2013; Zhou et al., 2021) held when 1) location/word pairs were presented simultaneously rather than sequentially and 2) errors due to intrusions from non target items were accounted for. In both cases, our findings corroborate the threshold account. [again, I’d reverse these to reflect their comparative theoretical importance.]

Firstly, although presenting location/word pairs simultaneously did decrease mean response error relative to the sequential format, the distribution of response errors were heavy tailed in both presentation formats. Despite the concern of Harlow and Donaldson (2013) that the simultaneous presentation of item and source would introduce a methodological confound in investigating source retrieval due to unitization between the two, we conclude that heavy tailed distributions of source errors, which has been interpreted as evidence of uniformly distributed guesses, is a methodologically robust finding.

Secondly, in Experiment 1, we found that intrusions accounted for some but not all error responses. While Bays et al. (2009) were able to eliminate the need for a uniform guessing component to account for errors in visual working memory, some proportion of high error responses in our source memory task appear not to be associated with non targets from the same study list. We found that a three-component model with both intrusions from non targets and guesses was strongly preferred over two-components model with either guessing or intrusions in isolation, suggesting that both of these processes contribute to error. Ultimately, our findings reinforce the position that source memory retrieval is thresholded and that participants guess when memory strength is subthreshold, though given our finding that intrusions account for some of these errors, prior estimates of the rates of guessing are likely overestimates (Harlow & Donaldson, 2013; Zhou et al., 2021). Rather than “a further contribution” I think this is the main thing, so again, lead with it and stress it. Two big contributions? First, to show, like VWM, that intrusions do contribute significantly to errors, but nevertheless the dual-process model continues to hold once intrusions are taken into account. Second, we’ve provided a detailed analysis of the spatiotemporal characteristics of the intrusion process. In VWM studies, in which intrusions have been identified as making a significant contribution to response error, items are presented simultaneously, but in source memory tasks, in which items are presented simultaneously, intrusions may be strongly affected by the serial position of targets relative to nontargets. We have shown that this is indeed the case, using a small*-N* design in which there were a large number of trials in each serial position, and found that both the spatial and the temporal proximity of nontargets to targets affected the probability of intrusions. These effects were both captured in a spatiotemporal gradient model in which the probability of an intrusion decreased exponentially with the spatial and the temporal distance between targets and nontargets.

A further contribution of our work was the introduction of similarity-based intrusion probability gradient models, which represents a novel attempt to systematically model similarity effects in memory in a continuous domain. Not only do we fit these models to response error, but we also account for distributions of response times, and demonstrate the value of the additional constraint such an approach affords.

### Theoretical Implications

Our work with the intrusion probability gradient models builds upon the Popov et al. (2021) finding that intrusions are more likely to come from adjacent lags than distant lags. Rather than separately estimate intrusions from different lags, we constrained our temporal gradient model by assuming that intrusions follow an exponential decay function with directional asymmetry. Finding temporal contiguity effects in these tasks is interesting because temporal similarity is not helpful in reporting the locations of words. Our findings support the assumptions of models like the temporal context model (TCM; Howard & Kahana, 2002), in which the forming of temporal associations is involuntary regardless of the task participants are presented with (Osth et al., 2019).

Further elaboration of the model with a spatial component that is multiplicatively combined with temporal similarity yielded the best fit to data. Unlike temporal information, the processing of spatial information was necessitated by our source memory task. Notably, Rerko et al. (2014) found a similar spatial transposition gradient when location was instead used as the cue to identify the item to be retrieved, where the retrieved information was the color of the item. From a visual working memory perspective, that we found clear support for spatiotemporal similarity in determining intrusion probability aligns with the interference model of Oberauer and Lin (2017) which argued that items are bound to context dimensions when studied, and that retrieval probability is given by the activation of retrieval candidates at test.

We did not find an effect of word feature similarity in the semantics or orthography of non targets, which was surprising given the body of work suggesting such features of words have strong effects in tasks of free recall and rates of false recall of orthographically similar or semantically related words (Conrad, 1963; Roediger & McDermott, 1995). Although the spatiotempotal context of non targets appears to dominate intrusion probabilities in the present work, our conclusions come with the aforementioned caveat that a replication of the current paradigm with word lists that are constructed specifically to maximize these kinds of similarity, such effects may yet manifest.

### Methodological Implications

As to whether intrusions are affected by similarity between the intruding item and the target, the various intrusion models were difficult to distinguish at the level of response errors in Experiment 1, resulting in a preference for the simplest three-component model when participants were fit at an individual level and then aggregated in terms of fit statistics. In this model all intrusions are equally probable irrespective of any kind of similarity, explaining its preference as the most parsimonious model when predictions between competing models are close. However, when decreasing the number of participants but increasing the number of trials performed by each participant in Experiment 2, we found that the spatiotemporal model was instead preferred for a majority of participants. Our interpretation of these seemingly conflicting results is that with the larger-*N* Experiment 1, there were insufficient observations at an individual level associated with intrusions responses to support the more sophisticated models of similarity-based intrusions, which was remedied by the small-*N* approach taken in Experiment 2. The methodological implications of this are expressed in Smith and Little (2018), specifically that increasing the sample size of participants would not have addressed the lack of power when considering the effect of similarity on intrusions at the individual level, and that by instead utilizing a small-*N* design we were able to concentrate power at the individual participant level.

While the choice of design speaks to where the quantity of data is concentrated, another difference highlighted in our results is the advantage of considering different types of data simultaneously. Even with the limitations of Experiment 1, when we considered the RT predictions of the diffusion models, we were able to more clearly differentiate between the models, which were constrained by the richer joint dataset. The clearest and most consistent evidence came from using both a small-*N* design and fitting RTs, that is, when comparing the diffusion models in Experiment 2. This adds to the growing body of work demonstrating that conclusions drawn on the basis of error data alone are less consistent, and in some cases later invalidated, by jointly modelling error and RT data (Ratcliff & Starns, 2009).

## Dynamic Rates of Guessing and Limitations of Mixture Modeling

One potential limitation of the family of models explored in the present study is we assumed that changes in the summed probability of intrusions across trials did not affect the probability of guesses, which remained constant. It may not always be reasonable to expect that the proportion of guesses remains the same across serial positions. To test this assumption in a coarse way, we implemented versions of the model where the parameter governing the proportion of guesses was separately estimated for the first and last items in the study list, but we did not find that these models made consistently different predictions from the base family of models and have therefore omitted them. To take another example, consider the potential interaction between recognition and intrusion probability where items that are not recognized do not intrude. In a list where no items are recognized, we would intuit that all responses should be guesses. A more rigorous approach requires a formal process model of how memory, intrusion, and guesses compete under different scenarios. This underscores the fundamental ambiguity of mixture models with more than two mixture components. A possible solution that could be explored in future work would be to implement the models introduced in this study in a race framework, such that the target and all non target responses are modelled as separate parallel evidence accumulation processes that compete to be retrieved in the manner in which discrete multi-alternative decisions[[5]](#footnote-6) have been modelled (Roe et al., 2001; Ratcliff & Starns, 2009; Leite & Ratcliff, 2010),which also compete with an additional process representing guesses as in the Timed Racing Diffusion Model (Hawkins & Heathcote, 2021). No, I don’t think so either. I think the theoretical alternative is clear enough: Each item in the list is associated with its own (circular) accumulator and the response is determined by the first one two finish.

Although several fruitful avenues for future research exist, the contributions of the present work represent a substantial theoretical advancement in systematically modelling similarity-based intrusions in a continuous domain. By developing and comparing this family of models across two datasets, we also demonstrate the utility of small-*N* designs and jointly modelling RT and accuracy data, both of which may prove useful to future work in the broader field of memory and decision-making research.

1. The von Mises distribution is a circular analogue of the Gaussian distribution. [↑](#footnote-ref-2)
2. We refer to erroneous responses driven by non target items in our task as intrusions, describing how words from non target word-location pairs are intruding on the cued pair. These within-list intrusions are not to be confused with extra-list intrusions, or *protrusion* errors, which we do not expect to contribute to errors in our paradigm. [↑](#footnote-ref-3)
3. The authors note that Healey (1974) observed a symmetric transposition gradient and Madigan (1971) observed asymmetry in the opposite direction, potentially due to differences in response modality. [↑](#footnote-ref-4)
4. Pretrained models were obtained from the fasttext.cc website, which were trained on the meta pages archive of English Wikipedia from June 2017, resulting in a text corpus of over 9 billion words in addition to news sources from statmt.org from 2007 - 2016, as described by Mikolov et al. (2017). [↑](#footnote-ref-5)
5. Multi-alternative decisions with continuous stimuli can be modeled with the circular diffusion model by partitioning the decision space with categorical boundaries (Smith, 2016). Similarly, the geometric framework of Kvam (2019) represents multiple alternatives as vectors in multidimensional space. In both of these cases, decisions are driven by a single evidence accumulation process towards a set of alternative response boundaries. This is distinct for our proposal, which describes a set of multiple accumulators that race in parallel (akin to earlier unpublished versions of the Ratcliff (2018) spatially continuous diffusion model (SCDM) to model the competition between target responses, intrusions, and guesses. [↑](#footnote-ref-6)