The Spatiotemporal Gradient of Intrusion Errors in Continuous Outcome Source Memory: Source Retrieval is Affected by both Guessing and Intrusions

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: This research was supported by Australian Research Council Discovery Grant DP210101787 to Philip L. Smith and Discovery Early Career Researcher Award DE170100106, awarded to Adam Osth.

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.

Commented [JZ1]: I really like the first part of Philip's title, but I think "implications for dual-process models" was missing the directionality of my original attempt. This one is a bit on the long side..

Commented [JZ2]: Check if this is still what should be listed for Adam

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 investigate whether these errors might instead reflect systematic intrusions from other list items which can mimic source guessing. Using the circular diffusion of decision making, which accounts for both response errors and RTs we found that intrusions account for some, but not all, errors in a continuous-report source memory task. Additionally, we found that intrusion errors were more likely to come from items studied in nearby locations and times, but not from semantically or perceptually similar cues. 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., Hautus et al., 2008; Osth et al., 2018; Onyper et al., 2008; Slotnick & Dodson, 2005; Yonelinas, 1999).

A key theoretical questions these models often address 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; Glanzer et al., 2004; Mickes et al., 2009). In contrast, threshold or discrete-state models assume that memory strength for an item must reach a certain threshold 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. 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 when it does not. 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).

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. In these tasks, 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.

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] 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. Subsequent work in visual working memory found that other sources of variability were required to explain distributions of response outcomes in VWM tasks. Bays et al. (2009) proposed an extension of the model in which on some trials, the incorrect item could be reported (an intrusion error). 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. Both variable precision and intrusion errors are critical features of the model we present below and have important implications for understanding the variability in continuous outcome responses.

Harlow and Donaldson (2013) introduced many of the same theoretical issues and experimental methodologies to long-term source memory with verbal stimuli. They used a continuous-outcome task in which word stimuli were paired with locations on the circumference

[1] The von Mises distribution is a circular analogue of the Gaussian distribution.

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[2]. On the basis that numerous studies have found a lack of source discriminability for unrecognized items (Malejka & Broder, 2016; Onyper et al., 2010; Bell et al., 2017; but see Fox & Osth, 2020 for exceptions), Zhou et al. (2021) conditioned source judgements on successful recognition and still found evidence for a threshold model. In the present study, we also condition source judgements on recognized items to distinguish between errors due to failures in item recognition and retrieval.

The Harlow and Donaldson (2013) interpretation attributes variability in response precision to two sources: 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. In the current paper, we consider two additional sources of variability: 1) the possibility that a nontarget item is reported instead of the target item and 2) variability due to properties of the decision-making process which acts upon the information retrieved from memory to generate the observed response. 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)

Commented [JZ3]: Added this bit to sneak in some reference to intrusions as early as possible.

[2] 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.

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. Previous work has attributed curvilinear zROCs to the joint contributions of familiarity and recollection. However, modeling of confidence judgments with the RTCON model (Ratcliff & Starns, 2009) -- an accumulator model that is jointly constrained by confidence and RT -- demonstrated that similar zROC functions could be produced with only a single source of evidence.

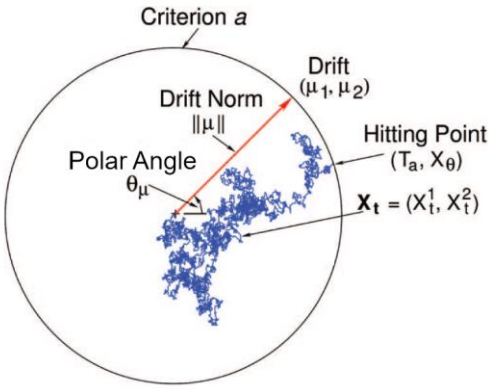
The diffusion decision model is a particularly influential account of decision making, which successfully explains well-documented phenomena like the speed-accuracy trade-off, and slow and fast error patterns under different decision conditions (Ratcliff et al., 2016). The diffusion model describes decision-making 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 vary 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*, 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



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, X_θ, is the decision outcome and the hitting time, T_a,is the decision time. The drift rate is vector-valued and consists of two components, (μ_1, μ_2), which jointly specify its magnitude and direction. In polar coordinates the magnitude is represented by the drift norm $\|\mu\|$ and direction is represented by the angle θ_μ 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, $\|\mu\|$, the decision criterion, *a*, and the noise in the evidence accumulation process, σ^2:

$$\kappa = \frac{a\|\mu\|}{\sigma^2} \quad (1)$$

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

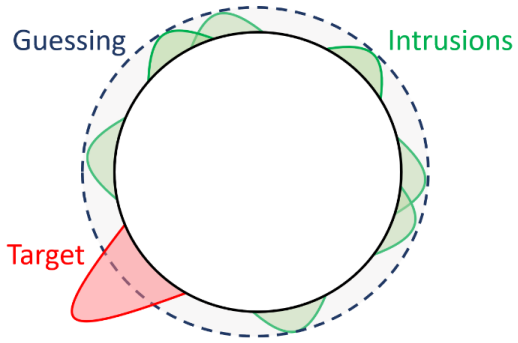
Through across-trial variability decision-making, specifically drift variability in the circular diffusion model, a single continuous process can produce distributions of response error with heavy-tails through the decision-making process, without invoking mixture with a uniform component in the memory 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, Zhou et al. (2021) 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.

As we noted previously, the VWM literature has highlighted the importance of distinguishing errors due to intrusions from errors due to guessing. Bays et al. (2009), 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. When the distribution of target features (in this instance, color) are random, response errors arising from confusions between a target and nontarget item will also appear to be uniformly distributed. Consequently, nontarget responses of this kind are indistinguishable from guessing when measured from the target. Figure 2 illustrates an example of how intrusions drawn from distributions centered on randomly dispersed nontargets on a single trial can mimic guesses from a uniform distribution across multiple trials.

Figure 2

Different Sources of Error in the Continuous-Outcome Task



Bays et al. (2009) noted that the two sources of error are differentiated by measuring the frequency of responses relative to all nontargets: guesses are uncorrelated with nontarget items, so the Zhang and Luck (2008) model 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 nontargets (Bays et al., 2009). That responses centered on nontarget items are more frequent than expected by chance is interpretable as evidence for nontarget responding.

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. Another 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 comparing models which make different assumptions about how intrusions arise.

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 followed by a centrally presented target word. Due to the temporal and spatial dissociation between item and source 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.

Intrusions

Most explanations of intrusions 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 nontargets are driven by confusions between items, then the probability of a given nontarget 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. The tendency for subjects to respond to nontarget 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[3], each reflecting specific properties of the tasks used to study the phenomenon (Bays, 2016). 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.

Sources of Intrusions

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, 2002a). In particular, participants are more likely to follow recall of an item with an item that was studied near it on the list. This tendency is more pronounced in the forward direction -- participants are more likely to follow recall of an item with an item from a later serial position than an earlier one (Kahana, 1996). Effects of temporal similarity have also been observed in transpositions errors in serial recall (Kahana & Caplan, 2002; Haberlandt et al., 2005; Farrell & Lewandowsky, 2004; Hurlstone & Hitch, 2014), as well as intrusion errors in paired-associate recall (Davis et al., 2008).

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,

[3] We refer to erroneous responses driven by nontarget items in our task as intrusions, describing how words from nontarget 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.

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 nontarget. Instead, 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 as Popov et al. (2021) did, 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 temporal distance (Howard & Kahana, 2002a; Murdock, 1997; Logan, 2021; Osth et al., 2018; 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) observed 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., 2022). 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 free recall, 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, 2002b; 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, 2002b).

Another feature that affects memory for words is their orthographic and/or phonological similarity (Conrad, 1963; Wickelgren, 1965). 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 with contextual factors.

Overview of Experiments in the Present Study

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 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 (Brysbaert & New, 2009). Word 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 library (de Leeuw, 2015) controlled stimulus presentation and recorded responses. 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.

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 $\pi/8$ 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 $\pi/8$ radians of the presented angle.

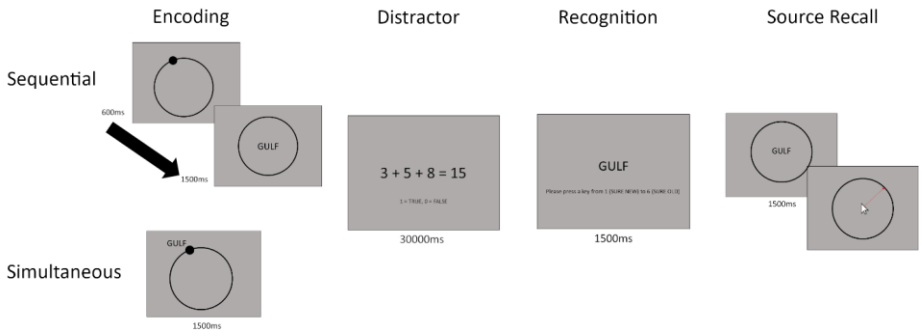
After studying each of the items for that block, participants were then instructed to complete a distractor task, which involved 30 seconds of arithmetic problems. These problems were presented as three single digit integers, which summed to a fourth number which would either be the correct sum, or a 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 center 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.



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 source judgments vary between sequential and simultaneous presentation of item and source information.

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 4
Normalized Histograms of Source Error in Sequential and Simultaneous Presentation Conditions Pooled across Subjects in each Condition

There was no significant difference between the mean absolute error for participants in the sequential (M = .06, SD = .04) and simultaneous (M = .08, SD = .05) presentation conditions $t(34) = 1.92$, $p = .063$. 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. 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

Our modeling 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. 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. We also implemented variations of some models that permitted different weightings for primacy and recency of intrusions and compared additive and multiplicative combinations of similarity. We have excluded these model variants in this text, but code for all the models developed is available at https://osf.io/76wtm/ and results are provided as supplementary material.

Model 1: Pure Guess

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

$$f(\theta;\kappa,\mu) = \frac{e^{\kappa\cos(\theta-\mu)}}{2\pi I_0(\kappa)} = \frac{e^{\kappa\cos(\theta-\mu)}}{\int_0^{2\pi} e^{\kappa\cos(\vartheta-\mu)}d\vartheta} \quad (2)$$

where the normalizing constant $I_0(\kappa)$ is a modified Bessel function of the first kind of order zero, which 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). 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

$$p(\hat{\theta}) = (1-\beta)\phi_\kappa(\hat{\theta}-\theta) + \beta\frac{1}{2\pi} \quad (3)$$

where the probability that a response is a guess is represented by β. In the target-driven component, θ represents the target angle, $\hat{\theta}$ is the reported angle, and ϕ_κ represents a von Mises distribution with a mean of 0 and with precision κ. 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 nontarget responses can be accounted for with intrusions from nontarget items without invoking guessing, model 2 substitutes the guessing component in the mixture model with an intrusion component:

$$p(\hat{\theta}) = (1-\gamma)\phi_\kappa(\hat{\theta}-\theta) + \gamma\frac{1}{m}\sum_{i}^{m}\phi_\kappa(\hat{\theta}-\theta_i) \quad (4)$$

where the probability of an intrusion occurring is represented by γ, and the angle associated with the i^{th} intruding item is represented by θ_i. Note that of the *m* nontarget items, the probability of a particular nontarget 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):

$$p(\hat{\theta}) = (1-\beta-\gamma)\phi_\kappa(\hat{\theta}-\theta) + \beta\frac{1}{2\pi} + \gamma\frac{1}{m}\sum_{i}^{m}\phi_\kappa(\hat{\theta}-\theta_i) \quad (5)$$

In contrast to subsequent models where the probability of a given nontarget 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 nontarget item intruding is determined by its temporal similarity to the target represented by *t:*

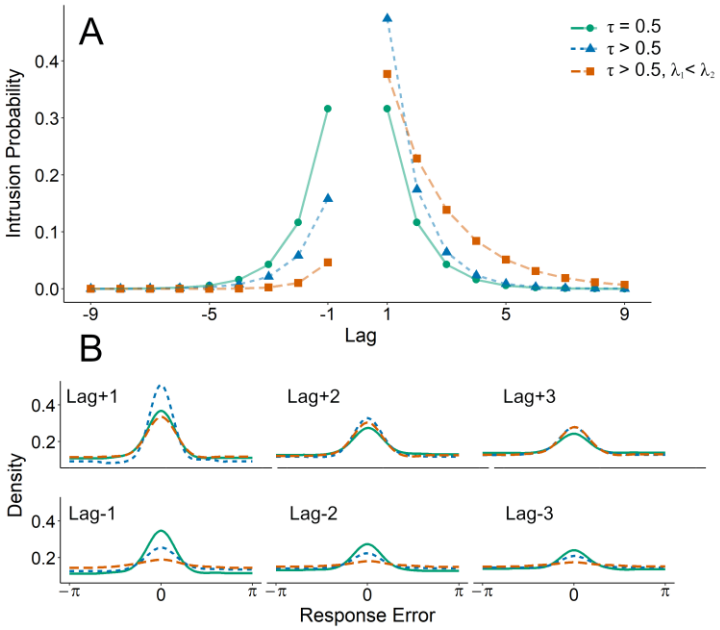
$$t = \begin{cases} \tau e^{-\lambda_1 l}, & l > 0 \\ (1-\tau)e^{-\lambda_2 l}, & l < 0 \end{cases} \quad (6)$$

We assume that the strength of association between items is an exponentially decreasing function of distance, a common property of models of contextual drift (Howard & Kahana, 2002a; Murdock, 1997; Logan, 2021; Osth et al., 2018). Temporal distance is expressed in terms of lag (*l*), which is defined as number of positions in the study list separating the target and

nontarget items. To allow for asymmetry in terms of temporal similarity for backwards and forwards lags, τ scales the similarity slope in each direction such that when $\tau > 0.5$, 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, λ, is estimated separately for the forwards and backwards similarity slopes. Figure 5 shows the effect of numerical variation in these parameters on model predictions.

Figure 5

Effect of Temporal Similarity Gradient on Model Predictions



Note. Panel A shows the shape of the temporal similarity gradient under various conditions: $\tau = 0.5$, $\lambda_1 = \lambda_2 = 1$ (solid green), $\tau = 0.7$, $\lambda_1 = \lambda_2 = 1$ (dotted blue), and $\tau = 0.7$, $\lambda_1 = 0.5$, $\lambda_2 = 1$ (dashed orange). Panel B shows the effect of these variations on simulated distributions of distances between response angles and nontarget angles. Each subpanel is conditioned on the lag of nontarget angles, such that greater central tendency is interpretable as a greater contribution of intrusions from each lag. Note that as λ increases, the predicted distribution of errors (relative to nontargets) approaches uniformity at a faster rate.

Commented [JZ4]: Might need to reword this for clarity, but struggling to do so right now.

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.

$$p(\hat{\theta}) = \left(1 - \beta - \gamma \sum_{i}^{m} t_i\right) \phi_\kappa(\hat{\theta} - \theta) + \beta \frac{1}{2\pi} + \gamma \sum_{i}^{m} t_i \phi_\kappa(\hat{\theta} - \theta_i) \quad (7)$$

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 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:

$$p(\hat{\theta}) = \left(1 - \beta - \gamma \sum_{i}^{m} w_i\right) \phi_\kappa(\hat{\theta} - \theta) + \beta \frac{1}{2\pi} + \gamma \sum_{i}^{m} w_i \phi_\kappa(\hat{\theta} - \theta_i) \quad (8)$$

$$w = t^{1-\rho} l^{\rho} \quad (9)$$

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*:

$$l = e^{-\zeta(1-\cos(\theta-\theta_i))} \quad (10)$$

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 nontarget item intruding is weighted by ρ. When $\rho = 0$, $w = t$ in Equation 9 and the weight function is independent of spatial location, while when $\rho = 1$, $w = l$ and the weight function is independent of lag. At intermediate values of ρ, w is jointly a function of the two. Naturally, intrusion responses from near nontargets will be associated with lower error relative to the target than intrusions from far nontargets. Therefore, as ρ 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 nontarget 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 single-character edits (insertions, deletions, or substitutions) to transform one string into another. Because all stimuli in our task were four-letter words, all edits were substitutions. We transformed raw Levenshtein distance into a measure of similarity scored by dividing each value by the four, the maximum number of edits, and subtracting the result from 1. The individual probability of a given nontarget item intruding is given by its weight, w:

$$w = (t^{1-\rho} l^{\rho})^{1-\chi} o^{\chi} \quad (11)$$

In words, the probability of a nontarget 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 nontarget and target word.

Model 7: Semantic Model

The semantic model substitutes orthographic similarity in model 6 for semantic similarity between target and nontarget words. To model semantic associations between words, 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]. 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; Mandera et al., 2016). Semantic similarity in our model, *s*, is defined as the cosine similarity between these vector representations, and is combined multiplicatively with spatiotemporal similarity to give the intrusion weight, *w*, for each nontarget item in each trial:

$$w = (t^{1-\rho} l^{\rho})^{1-\chi} s^{\chi} \quad (12)$$

Model 8: Four-Factor Model

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:

[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).

$$w = (t^{1-\rho} l^{\rho})^{1-\chi} (o^{1-\psi} s^{\psi})^{\chi} \quad (13)$$

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

Table 1

Summary of Response Error Models and Parameters

Parameter	Description
κ_1	Precision, memory
κ_2	Precision, intrusion
β	Proportion of uniform guesses
γ	Proportion of intrusion responses
τ	Temporal gradient asymmetry
λ_1	Temporal similarity decay, forwards
λ_2	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	κ_1, β	2
2. Pure Intrusion	$\kappa_1, \kappa_2, \gamma$	3
3. Intrusion + Guess (Flat)	$\kappa_1, \kappa_2, \beta, \gamma$	4
4. Temporal	$\kappa_1, \kappa_2, \beta, \gamma, \tau, \lambda_1, \lambda_2$	7
5. Spatiotemporal	$\kappa_1, \kappa_2, \beta, \gamma, \tau, \lambda_1, \lambda_2, \zeta, \rho$	9
6. Spatiotemporal-Orthographic	$\kappa_1, \kappa_2, \beta, \gamma, \tau, \lambda_1, \lambda_2, \zeta, \rho, \chi$	10
7. Spatiotemporal-Semantic	$\kappa_1, \kappa_2, \beta, \gamma, \tau, \lambda_1, \lambda_2, \zeta, \rho, \chi$	10

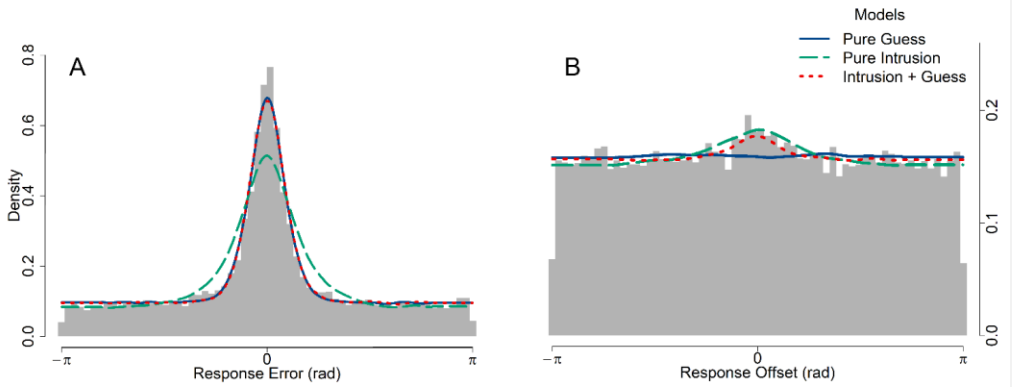
8. Four-Factor	$\kappa_1, \kappa_2, \beta, \gamma, \tau, \lambda_1, \lambda_2, \zeta, \rho, \chi, \psi$	11

Response Error Model Comparison

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 nontarget 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 6B. We will subsequently refer to this analysis as *recentering* the data, as it is equivalent to recentering the distribution of response errors on the nontarget angles.

Figure 6

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



Note. Observed data are represented by gray histograms, while model predictions are represented by dashed lines. 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. The error

predictions of the pure guess and intrusion + guess models are indistinguishable; the predicted distributions lie on top of each other in panel A. In panel B, distances are instead calculated between the response angle and each nontarget angle, that is, the location of all other items in the block excluding the trial target.

While model 2 (pure intrusion) is able to predict heavy tails in the distribution of response errors (Figure 6A), it underpredicts the height of the central peak of the distribution, that is, the precision of memory responses. In contrast, model 1 (pure guess) provides a good account of the distribution of response errors, but misses the central tendency evident in the recentered data shown in Figure 6B because it does not predict any relationship between the response and nontarget angles. Model 3, with both guessing and intrusion components, is able to simultaneously produce both patterns of data, suggesting that both processes are needed to explain the distribution of response angles relative to target and nontarget angles. To further illustrate this point, Table 2 shows the best fitting parameter values for each model averaged across participants. With specific reference to the estimated proportion of guesses, 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 nontarget responses ($\beta = 0.60$ in the former, $\beta + \gamma \approx 0.60$ in the latter).

Table 2

Average Parameter Estimates for Each Model to Experiment 1 Data.

Model	Parameter Average										
	κ_1	κ_2	$\boldsymbol{\beta}$	γ	τ	λ_1	λ_2	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		

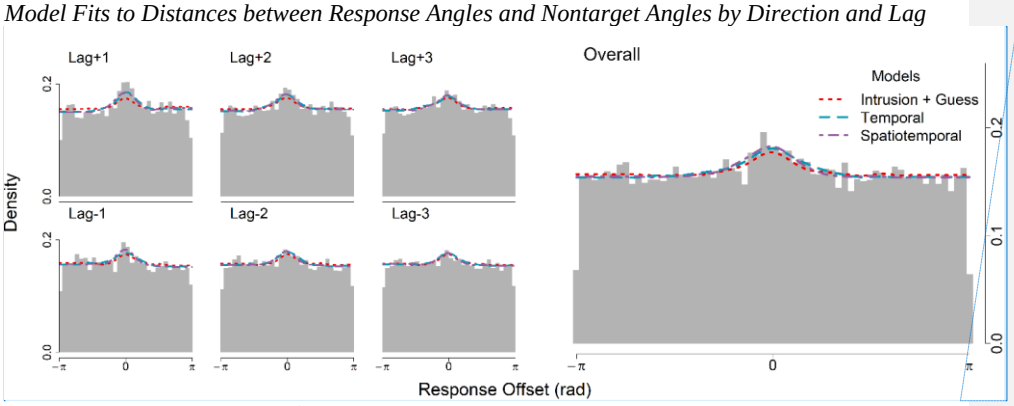
6	14.66	12.67	**0.49**	0.42	0.59	1.97	2.18	0.53	0.40	0.34	
7	16.84	11.27	**0.54**	0.19	0.59	1.64	1.50	0.39	0.46	0.50	
8	14.21	12.33	**0.51**	0.46	0.45	1.82	2.02	0.51	0.29	0.34	0.24

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 compare models in which spatial, temporal, semantic, and orthographic similarity between targets and nontargets affected the intrusion probability. Models 3 (intrusion + guess), 4 (temporal), and 5 (spatiotemporal) make visually indistinguishable predictions about the overall distribution of response errors as well as the overall recentered errors (Figure 7). Instead, the effect of different intrusion probability gradients can be seen by conditioning the recentered data on the lag and direction of each intrusion. Central tendency, or the contribution of 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 model assumes that intrusions are equally likely from all nontarget 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 7). We also fit further elaborations to the intrusion component of the spatiotemporal model (that is, 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 7 and 8 to emphasize the differences between the intrusion + guess, temporal, and spatiotemporal models.

Figure 7

Model Fits to Distances between Response Angles and Nontarget Angles by Direction and Lag



Note. Each subpanel on the left shows errors relative to nontarget angles, conditioned on the intrusion lag, or the number of positions separating each nontarget angle and the target angle. The larger panel on the right shows the overall distribution of intrusion offsets collapsed across lags, identical to Figure 6B. While the spatiotemporal and temporal gradient models are difficult to distinguish from the flat intrusion + guess in the overall plot, the gradient models are better able to produce the changes in central tendency with lag direction and magnitude visible in the conditional subpanels.

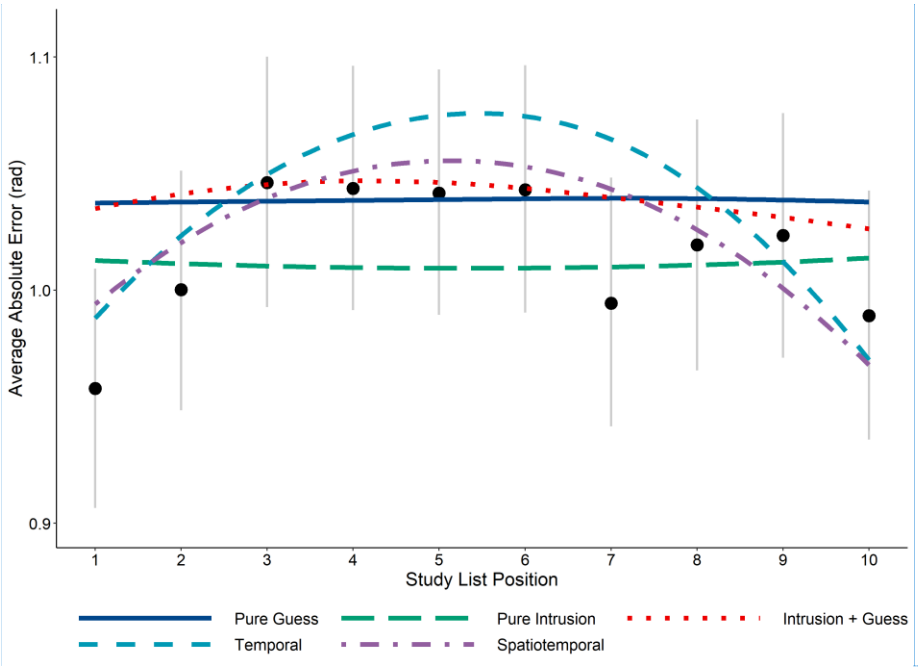
Commented [JZ5]: This figure is probably still a bit hard to read. Considering just excluding the spatiotemporal gradient model as well, just to highlight the temporal gradient vs no gradient, maybe stretching the y-axis for each subplot.

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 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 temporally similar nontargets. 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. However, 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 nontargets 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



Commented [JZ6]: I think thickening the lines and choosing different line types has made this figure (and most of the others) clearer. Looks ok in b&w too. Thanks Philip.

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 the qualitative advantages of the two gradient models over the intrusions + guess model, the latter is preferred in a quantitative sense on the basis of the Akaike information criterion (AIC). Models were fit to 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 fit of each model relative to the best fitting model, expressed

in terms of the difference between AIC values summed across participants (ΔΣAIC), such that the best fitting model has a value of 0. These 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 temporal and temporal gradient models. Individual *w*(AIC) comparisons are provided as supplementary material. 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
2. Pure Intrusion	3	38178.07	1116.16	0
3. Intrusion + Guess (Flat)	**4**	**37061.91**	**0**	**1**
4. Temporal Gradient	7	37176.82	114.91	0
5. Spatiotemporal Gradient	9	37237.68	175.77	0
6. Orthographic	10	37633.28	569.28	0
7. Semantic	10	37310.81	246.81	0
8. Four-Factor	11	37705.13	641.13	0

The serial position dependence of response errors in Figure 8 favors a spatiotemporal gradient, but the best model in an AIC sense (Table 3) was Model 3 (intrusions + Guess) 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 (Ratcliff & Starns, 2009). 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 circular diffusion models we employ resemble the above models, but require an additional parameters for the decision criterion (the point at which enough evidence has been accumulated to reach a decision) and the standard deviation of the drift rate, which is analogous to variability in precision. 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 nontarget responses, and so these parameters were estimated separately for the memory component (μ_1, η_1) and the intrusion component (μ_2, η_2), 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.

In difficult two-choice decisions in which accuracy is stressed, error RTs are typically slower than correct RTs (Luce, 1986, p. 233). A property of the circular diffusion model, inherited from the diffusion model of two-choice decisions (Ratcliff, 1978; Ratcliff & McKoon, 2008) is that it can predict this phenomenon, known as a slow-error pattern. 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, we not only assume that drift rates vary across trials, but also that intrusion responses are 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 we present. 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 (T_{er}) 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	μ_1, η_1, a_1, a_2, T_{er}, β **(6)**	47611.67	1819.00	0
2. Pure Intrusion	μ_1, η_1, μ_2, η_2, a_1, T_{er}, γ **(7)**	46512.06	719.39	0
3. Intrusion + Guess (Flat)	μ_1, η_1, μ_2, η_2, a_1, a_2, T_{er}, β, γ **(9)**	45850.07	57.41	0
4. Temporal	μ_1, η_1, μ_2, η_2, a_1, a_2, T_{er}, β, γ, τ, λ_1, λ_2 **(12)**	45988.75	196.09	0

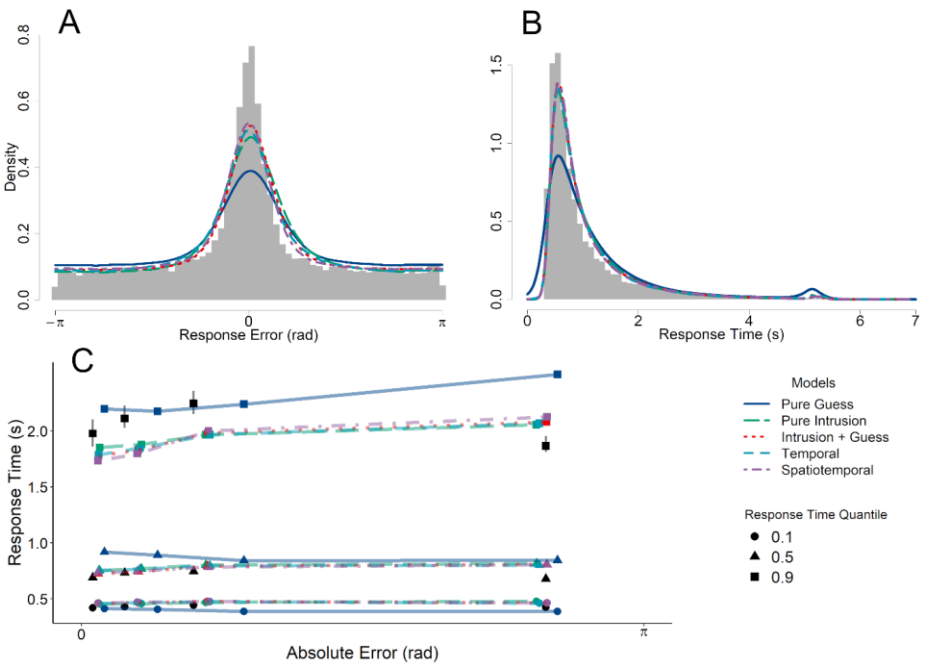
5. Spatiotemporal	μ_1, η_1, μ_2, η_2, a_1, a_2, T_{er}, β, γ, τ, λ_1, λ_2, ζ, ρ **(14)**	45792.67	0.00	1

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. Figure 9A 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 7A, 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 9A and 9B respectively, Figure 9C 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 9C, 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.

Commented [7]: I think this is a nice demonstration of the value-add of the diffusion model over the simpler marginal error model. The circular diffusion model predicts more structure (joint distributions) and it favors models that can predict this structure. These effects are invisible to the marginal error models.

Figure 9

Diffusion Model Fits to Response Error and Latency



Note. Grey lines in Panel C 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	μ_1	μ_2	η_1	η_2	a_1	a_2	γ	β	τ	λ_1	λ_2	ζ	ρ	T_{er}
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, we have found clear evidence that intrusions from nontargets 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 (Harlow & Donaldson, 2013; Zhou et al., 2021). However, the poor fit of the pure intrusion models, again both in terms of error and joint error and RT data, suggests that a purely continuous view of source memory retrieval is incompatible with the data, even when intrusions are accounted for.

Secondly, and contrary to our expectations, elaborations of the intrusion component to model the effect of similarity on intrusion probability 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 nontarget 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. We also found, when recentering response errors on nontarget

angles, we found clear evidence for temporal intrusions. 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).

Finally, 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 unaltered.

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

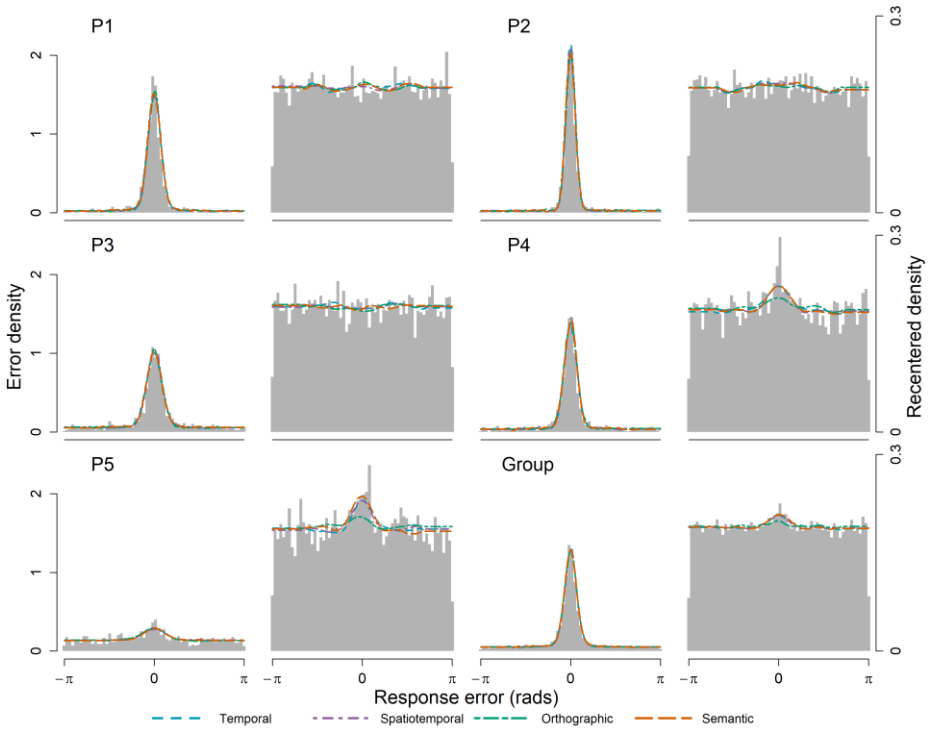
Source and item information was only presented simultaneously for all participants in Experiment 2. To maximize the number of trials, in turn to maximize the diagnosticity of our model selection procedure, we presented source and item information simultaneously on all trials of the experiment.

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 this section, we compare in more detail models in which the intrusion probabilities also depended on the orthographic and semantic similarity between targets and nontargets. Figure 10 shows the graphical fits of the models to each participant-level dataset, both in terms of response error and errors recentered on nontargets. Recentered plots conditioned on levels of orthographic and semantic similarity were not diagnostic and have been omitted and are instead provided as supplementary material.

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 10 are difficult to distinguish. One concern in comparing models is the diagnosticity of the results 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 explicitly manipulated when study lists were constructed and the average similarity between items was consequently low. We elaborate this point in the discussion section to follow. 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 5).

Table 5

Parameter estimates for each model, averaged across participants

Model	Parameter Average										
	κ_1	κ_2	β	γ	τ	λ_1	λ_2	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	21.93	15.88	0.30	0.20	0.75	1.89	1.55	0.88	0.34	0.50	0.01

In Table 6, which quantitatively compares the models an individual level, Model 5 (spatiotemporal) is preferred for majority of participants to varying degrees. 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, models 4 (temporal) and 6 (orthographic) are more competitive. The orthographic model is preferred outright for Participant 2, while for

Participant 5 the simpler model 3 (intrusions + guess) 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
			AIC			
1	1116	1112	**1088**	1095	1096	1097
2	555	558	562	**550**	563	559
3	2363	2364	**2361**	2367	2368	2371
4	1520	1499	**1498**	1498	1501	1577
5	**3557**	3567	3566	3567	3568	3612
			w(AIC)			
1	0	0	**0.94**	0.03	0.02	0.01
2	0.08	0.02	0	**0.82**	0	0.07
3	0.23	0.14	**0.57**	0.03	0.02	0.01
4	0	0.25	**0.39**	0.28	0.08	0
5	**0.96**	0.01	0.02	0.01	0.01	0

Diffusion Models

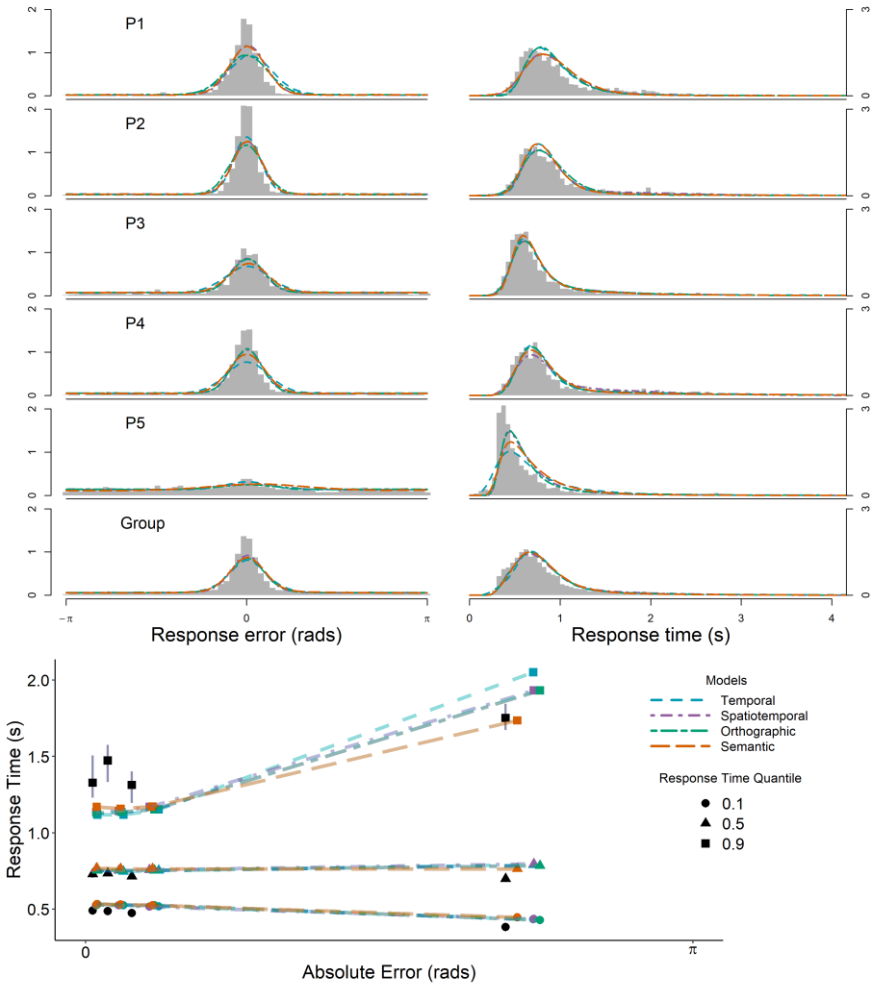
As with Experiment 1, we also compared circular diffusion versions of each model. The graphical fit of the models to response error and times is shown in Figure 11. Model 8 (four-factor) performed worse than the three-factor semantic and orthographic models and has been omitted from Figure 11 to better present the other model predictions. As with Experiment 1, the diffusion model appears to misfit the distribution of response errors to a greater degree than the

Commented [8]: Maybe need a table that establishes a correspondence between the names of the models and their numbers then? It's hard to infer from the lists of parameters in Table 5. And when you refer to them in text, refer to them both by name and number.

models presented in Figure 10. This is because the diffusion models attempt to fit joint RT distribution and accuracy data, whereas the earlier analysis fits only the error data, marginalized over RT. 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 and Group Level Response Error and Response Times



The joint relationship between response error and RT is most clearly demonstrated at a group level, shown in the bottom panel of Figure 11, which plots response time quantiles for the data binned by response error quantiles. The spatiotemporal gradient model makes closer predictions than the temporal model, which tends to overpredict the magnitude of the slow error

pattern. The semantic and orthographic components, when added to the spatiotemporal model, do not substantially improve the model fit, and the predictions of these models lie on top of each other when plotted. Notably, all models under consideration misfit the .9 RT quantiles for the three most accurate error bins. This facet of the data can be interpreted as a proportion of accurate responses which are slower than the models predict. 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.

Table 7

Average Parameter Estimates

Model	Average Parameter Estimates							
	μ_1	μ_2	η_1	η_2	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	4.26	0.16	0.11	0.01	2.72	1.29	0.09	0.34
	τ	λ_1	λ_2	ζ	ρ	χ	ψ	T_{er}
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	0.74	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. Compared to the models fit to response error data, the results of the diffusion model comparison provide more stable support for the spatiotemporal model, owing to the additional constraints imposed on the models by jointly fitting the RT data.

Table 8

Experiment 2 Diffusion Model AIC Comparison

Participant	Model					
	3	4	5	6	7	8
			AIC			
1	2289	2333	**2126**	2347	2153	2347
2	2259	2053	**2032**	2274	2171	2104
3	3494	3330	**3240**	3243	3299	3306
4	3425	3266	**2892**	3069	3096	3012
5	4260	4469	**4054**	4069	4254	4142
			w(AIC)			
1	0	0	**1**	0	0	0
2	0	0	**1**	0	0	0
3	0	0	**0.78**	0.22	0	0
4	0	0	**1**	0	0	0
5	0	0	**1**	0	0	0

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) errors due to intrusions from nontargets are distinguished from “memory-less” guessing, and 2) when location/word pairs were presented simultaneously rather than sequentially. In both cases, our findings support a thresholded view of source memory retrieval.

Intrusions

We found that intrusions do contribute significantly to errors, and once these intrusions are explicitly accounted for, the proportion of responses attributable to source guessing dropped accordingly. However, unlike Bays et al. (2009), who were able to eliminate the need for uniform guesses to account for errors in VWM, a proportion of high error responses in our source memory task could not be explained by intrusions. Instead, we found that a three-component model with both intrusions and guesses was strongly preferred over two-component variants with guessing or intrusions exclusively. Ultimately, our findings suggest that although previous studies have overestimated the proportion of source retrieval failures, such failures do occur, which is consistent with the predictions of a source retrieval threshold and of the dual-process model (Harlow & Donaldson, 2013; Yonelinas, 1999).

Contextual Similarity

We provide a detailed analysis of the similarity-based characteristics of the intrusion process. Specifically, we found that the temporal and spatial proximity of nontargets to targets at study affected the probability of intrusions. The temporal component of our model 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, 2002a), in which the forming of temporal associations is involuntary regardless of the task participants are presented with (Osth et al., 2019).

By combining temporal similarity with spatial similarity in the spatiotemporal gradient model, we draw an explicit link between interference effects in VWM. In VWM studies, in which intrusions have been identified as making a significant contribution to response error, items in a set are presented simultaneously, but in source memory tasks, in which items are presented as a list, intrusions may be strongly affected by the serial position of targets relative to nontargets (Oberauer & Lin, 2017; Rerko et al., 2014). 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 temporal distance between targets and nontargets. Our study joins a growing body of work that suggests that an item's position in time and space are not uniquely privileged features, but fundamentally similar contextual dimensions along which items in memory differ, and that confusions between items along these dimensions are governed by similar gradients of errors (Oberauer & Lin, 2017; Schneegans et al., 2022).

Item Similarity

Contrary to our expectation that the similarity-based intrusion component in our model would be further improved by adding item-based similarity to the model with reference to the semantics and orthography of the word stimuli, we did not find an advantage when comparing these models to the spatiotemporal model. One possible explanation for this null result 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, in our dataset a Levenshtein distance of 1 (an equivalent orthographic measure occurred on less than 3% of the presented lists. However, it is worth noting that semantic similarity can exert effects even in lists of unrelated words, including transitions between list words in a free recall task (Howard & Kahana, 2002b; Morton & Polyn, 2016) and predicting false alarms in recognition memory (Osth et al., 2020).

That we did not observe an effect of semantics may also be due to the particular demands of the source task. When items are presented individually on a study list, items are associated to the list context and to other items on the list (e.g., Gillund & Shiffrin, 1984). Semantic similarity can exert a large effect on recall transitions because each recalled item is used as a cue for further retrievals – semantically similar items facilitate this process. In recognition memory, items are matched against all of the other items on the list to produce an index of global similarity that is the basis of the recognition decision (e.g., global matching: Clark & Gronlund, 1996; Osth & Dennis, 2022). Thus, semantically similar items on the list will contribute to the global similarity and increase the likelihood of a false alarm. In a source task, in contrast, items are associated to the source of their occurrence, and it is not necessarily beneficial to associate items to other list items. At retrieval, both the list context and the item cue are used to retrieve the specific source location. Given the lack of associations formed between items, semantic similarity between the items may exert less of an influence than in a task such as free recall.

A very likely possibility is that both semantic and orthographic similarity do exert an influence on intrusion probabilities, but the between-word similarities in our dataset were too

low to exert a noticeable influence due to the reasons listed above. The fact that spatiotemporal similarity exerted a dominant role may simply be because the levels of both temporal and spatial location similarity were comparatively much higher. Virtually all existing models would predict effects of both word similarity and spatiotemporal similarity, including the interference model of continuous report (Oberauer & Lin, 2017), but also models of episodic memory, in which retrieval is heavily influenced by the similarity of both item and context representations (e.g., Cox & Shiffrin, 2017; Gillund & Shiffrin, 1984; Murdock, 1997; Osth & Dennis, 2015; Polyn, Norman, & Kahana, 2009). An effect of semantic and orthographic similarity on intrusion probability may be observed in a future replication of the current paradigm with word lists that are constructed specifically to maximize similarity along these dimensions.

Estimating Proportions of Multiple Components in Mixture Models

One limitation of the family of models explored in the present study is that 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 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 each item in the list associated with its own accumulator, which compete to be retrieved in the manner in which discrete multi-alternative decisions[5] have been modelled (Roe et al., 2001; Ratcliff & Starns, 2009; Leite & Ratcliff, 2010). To model intrusion effects in conjunction with guesses, accumulators for each item would also compete with an additional process representing guesses as in the Timed Racing Diffusion Model (Hawkins & Heathcote, 2021).

In conclusion, we found that both intrusion errors and guesses, due to source retrieval failures, explain response error data in continuous outcome source memory tasks. We argue that previous research has overestimated the proportion of guesses by conflating these two source of error. We arrive at this conclusion by modeling both RT and accuracy with the circular diffusion model, which allows identification of independent sources of variability from memory and decision processes. By distinguishing both intrusion errors from guesses, as well as properties of memory from decision-making, we provide principled evidence in support of the view that source retrieval is a thresholded process.

[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 different from 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.