Abstract

Intrusions occur when information about a non-target item is erroneously reported in place of the target item and have been observed in tasks ranging from serial recall to visual working memory tasks. Using the circular diffusion of decision making we found that intrusions account for some, but not all, errors in a continuous-report source memory task. Additionally, we found that the probability of a given non-target intruding is determined by the similarity of the spatiotemporal context in which it and the target item are presented, and not the semantic or orthographic similarity of the words.

*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 a 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. Several models have been advanced to understand the processes governing both recognition and source judgements (Yonelinas, 1999; Slotnick & Dodson, 2005; Hautus et al., 2008).   
 A key question such models contend with is whether the retrieval of information from source memory is better characterized as a continuous or a discrete process. Models of memory retrieval as a continuous process, based upon Signal Detection Theory (SDT), assume that memory strength varies continuously, and so predict that performance in a source memory task declines gradually as memory strength decreases (Banks, 2000; Mickes et al., 2009). In contrast, threshold or discrete-state models assume that memory strength for an item must reach a certain threshold in order for that item to be retrieved, and so predict that source responses are either made with high precision when driven by memory or are guesses, made in the absence of information, when the memory is below the retrieval threshold (Batchelder & Riefer, 1990; Klauer & Kellen, 2010). Another alternative is the dual-process framework, in which different retrieval mechanisms are used in different kinds of memory tasks (Mander, 1980). Specifically, the two processes in the influential Yonelinas (1999) dual-process model are 1) familiarity, which yields a continuous measure of strength for an item in memory and 2) recollection, which yields richer contextual information about the study event through a search process which is thresholded. Successful recollection or familiarity can both contribute to recognition, because familiarity can distinguish between a studied and an unstudied item. On the other hand, familiarity does not distinguish between two items from different sources, which are both studied, and so the Yonelinas (1999) dual-process model predicts that source judgements should be thresholded as they can only be driven by recollection. This dual-process view of memory retrieval holds only if recollection, and therefore source memory performance, can be characterized as a thresholded process. Existing research which attempted to distinguish between continuous and thresholded models of source memory has been based on data from two-choice tasks, whereby confidence ratings and accuracy in two-choice tasks are used 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 found numerous conditions under which the models mimic each other (Yonelinas & Parks, 2007; Klauer & Kellen, 2010).

**Continuous-Outcome Tasks**

An alternative to using two-choice tasks is to use *continuous-outcome* tasks, in which responses are made on a continuous scale. The advantage of using such a task is that it allows direct measurement of response precision, as opposed to the proportion of responses in each of the discrete options in a traditional two-choice task. 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). Continuous-outcome tasks were first used to study memory in the specific context of how visual working memory (VWM) representations change with the number of items stored in memory (Wilken & Ma, 2004). 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. In both cases, the common question about the architecture of memory is if information is stored in discrete states. Zhang and Luck (2008) modelled distributions of response outcomes in a color recall task under different set size conditions, and found the data was well described by a mixture model, specifically a mixture of a von Mises distribution[[1]](#footnote-1) and a uniform distribution.

Applying a similar approach to source memory modelling, Harlow and Donaldson (2013) used a continuous-outcome task in which source was operationalized as the locations of word stimuli along the circumference of a circle. At test, participants reproduced remembered locations on a response circle when cued with each word in the study list. The authors found that a mixture model consisting of a wrapped Cauchy and a uniform component was preferred over a pure wrapped Cauchy model, which was interpreted as evidence for a thresholded retrieval process which yields uniform guesses when memory strength is subthreshold (Harlow & Donaldson, 2013). This analysis attributes variability in response precision to two sources in memory: 1) variability in memory precision and 2) the possibility that memory is absent and the response is a guess. To account for the role of decision-making processes in generating responses, Zhou et al. (2021) applied the circular diffusion model, a model of decision-making in circular domains, which decomposes variability into that arising from memory and decision-making processes.

### Decision-Making in Continuous-Outcome Tasks

Any response observed in a memory task is a product of a decision-making process in addition to the information from memory driving the decision. Accurately characterizing the effect of decision-making is critical to understanding the nature of memory retrieval (Ratcliff, 1978). Accounting for decision-making over time requires analysis of not only response outcome, but also response time (RT) data, and considering both types of data yields more diagnostic information about the underlying cognitive processes. The importance of modeling decision-making is well illustrated in the recognition memory literature, where initial conclusions founded on the shape of ROC curves were later challenged by including RT data in addition to the response proportions used to form ROCs (Ratcliff & Starns, 2009; Starns et al., 2012; Dube et al., 2013; Osth et al., 2017), and in characterizing serial position effect in free recall (Osth & Farrell, 2019).

A particularly influential account of decision-making is the diffusion decision model, 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).

The circular diffusion model inherits the desirable explanatory qualities of the standard two-choice diffusion model and extends the model to a continuous response space by representing evidence accumulation as a vector in two-dimensional space that starts at the origin of a circle and terminates at a point in its circumference, which represents the decision outcome (Smith, 2016). The introduction of the circular diffusion model addressed the lack of a formal model of RT and decision-making in continuous-outcome tasks, enabling use of the noted advantages of the paradigm over two-choice tasks in addition to RT modeling.

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

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

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

Through across-trial variability decision-making, specifically drift variability in the circular diffusion model, a single continuous process can produce distributions of response error with heavy-tails through the decision-making process, without invoking mixture with a uniform component in the memory process (van den Berg et al., 2012; Smith, 2016). By jointly fitting response error and response time (RT) and error data, Zhou et al. (2021) found that thresholded model with a uniform component was preferred over a continuous model with drift variability, broadly corroborating the conclusions of Harlow and Donaldson (2013).

## Non-target Responding

In the VWM literature, the slots account of memory capacity proposed by Zhang and Luck (2008) is built upon the finding that a proportion of responses appear to be uniformly distributed and reflect random guessing. Bays et al. (2009) challenged this interpretation by arguing that confusions between target and non-target items could also account for errors that appear uniform relative to the target item. The authors demonstrated that a model incorporating probability distributions centered on the identity of each non-target item in the display accounted for the distribution of response errors without a discrete limit on the number of items stored (this model is formally described below). Confusions between items, such that information about a non-target item is reported in place of the cued target, result in *swap errors*, and because the possibility of these swaps can be confounded for variability in memory precision, disentangling these sources of error has been important in accurately characterizing VWM processes (Bays, 2016; Rajsic & Wilson, 2012, 2014; Pertzov et al., 2012). Applying this approach to source memory, we investigate the extent to which non-target responses, caused by intrusions between item-source pairs, account for source errors.

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

### Contiguity Effects

The tendency for subjects to respond to non-target features or items has been observed in a wide variety of cognitive tasks, and the related types of errors that arise are referred to by various terms including *binding, transposition, intrusion,* and *swap errors*[[2]](#footnote-2), each reflecting specific properties of the tasks used to study the phenomenon (Bays, 2016). Most explanations of non-target responding attribute the phenomenon to confusion between items that are similar (Rerko et al., 2014; Bays, 2016; Oberauer & Lin, 2017; but see Pratte, 2018 for an alternative view). 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 distance in the location of the items, as well as features of the words used as stimuli such as their semantic and orthographic similarity, the effects of which have been studied across the broader body of episodic memory research. In the paragraphs to follow, we review the commonalities between findings across different tasks, all of which motivate the present modelling of source memory.

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). In free-recall tasks, where participants are asked to recall a list of items in any sequence they wish, Kahana (1996) demonstrated that after recalling a given item, the next item to be recalled tends to be a neighboring item in the study sequence. The distance between item *i* in the study list and another item in the list is known as the *lag*. Additionally, neighbors in the forwards direction (*i* + lag) were more likely to follow an item than backwards neighbors (*i* – lag), referred to as forward asymmetry in the contiguity effect. The probability of transitioning to a given lag at recall is known as the lag-conditional response probability (lag-CRP), and the effect of increasing lag on this probability can be seen in Figure 1.

Figure 1

*Transition gradient seen in lag-CRP in free recall*

Chart

Description automatically generated

While associations between temporally contiguous items can facilitate responses in free-recall tasks, the same type of association can contribute to errors in tasks when the sequence of items is important. Specifically, in serial recall tasks, when subjects must call lists of items in the sequence in which they are given, a classic finding is that incorrect responses tend to be items studied near the target in the study sequence (Lee & Estes, 1977; Nairne, 1990). This effect can be described as a transposition of two items, in that the position of non-target items are swapped with that of the target item, and as in free-recall, the probability of a swap is inversely related to the distance of the two items (Henson et al., 1996; Page & Norris, 1998; Lewandowsky & Farrell, 2008). While early models explained this effect through a “chaining” process by which items were bound to each other in sequence, more recent explanations argue that items are instead associated with representations of their serial position, and that these representations overlap so that an item can cue not only its position but that of its neighbors (Lewandowsky & Farrell, 2008; Rerko et al., 2014). Applying the lag-CRP methodology to serial recall data forms a *transposition gradient* around the target location effect (Kahana & Caplan, 2002; Solway et al., 2012). Like the shape of the lag-CRP curve in free recall, transposition gradients tend to exhibit a forwards asymmetry both in terms of transposition probability (Klein et al., 2005; Haberlandt et al., 2005[[3]](#footnote-3)) as well as latency (Farrell & Lewandowsky, 2004; Hurlstone & Hitch, 2014).

The temporal contiguity effect has also been observed in paired-associate recall. After studying pairs of words, Davis et al. (2008) found that when participants recalled non-target items, the erroneous item tended to be intrusions from temporally contiguous pairs. One again following the logic of lag-CRP analyses, the probability of an intrusion from pair *i* + lag when cued with an item from pair *i* decreases with absolute lag, and is asymmetric in the forwards direction so that intrusion probability is greatest when lag = 1.

More 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 the differences in accuracy for high-frequency and low-frequency words was attributable to higher levels of binding errors for low-frequency words and not memory precision of the proportion of guesses. In particular, when participants made a mis-binding error, responses were not generated from a random non-target. As with intrusions from paired words (Davis et al., 2008), Popov et al. (2021) found that mis-binding errors were most likely to come from locations in neighboring serial positions by separately estimating the probability that a response came from each of the locations in the study set. The authors demonstrated a contiguity effect by comparing the probability of mis-binding across lags (Popov et al., 2021). The present study aims build upon this body of work by systematically modelling the rate at which intrusion probability decreases with lag.

In the same way that temporal contiguity effect describes how limitations of temporal distinctiveness explains transition and transposition gradients in memory for lists of items, Rerko et al. (2014) refer to an analogous effect in the spatial domain to explain similarly graded effects of distance, in that spatial confusions between items are more common at smaller distances (Emrich & Ferber, 2012; Bays, 2016; Sahan et al., 2019). The link between swap errors in VWM and transposition errors in serial recall has been proposed to reflect a more general mechanism in memory by which items are bound to context dimensions (Oberauer & Lin, 2017; Schneegan et al., in press).

### Source Memory for Associated Items

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

Another feature that affects memory for words is their orthographic and/or phonological similarity (Conrad, 1963; Wickelgren, 1965). The effect of these features on false memory specifically has also been studied using the DRM paradigm. Sommers and Lewis (1999) constructed lists of phonologically related words and found that rates of false recall were highest when words were close phonological neighbors and lower when words were phonologically dissimilar, suggesting that words that differ by a single grapheme or letter were most likely to be confused. When words are presented visually, prior research about reading suggests that both orthographic and phonological features are activated and encoded in memory (Massaro & Cohen, 1994). In contrast, when words are presented aurally, orthographic features of words are thought to decay more rapidly than phonological features (Tannenhaus et al., 1980). When comparing written and spoken presentation modalities, Smith and Hunt (1998) found that levels of false memory were around twice as high for spoken words than for written words (on the modality effect, also see Kellogg, 2001). 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 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)

## The Present Study

In the present study, our primary aim was to investigate systematic models of intrusions in the continuous-outcome source memory task, using models of both response error and response times. Over the course of two experiments, we compare models in which intrusion probabilities are sensitive to temporal similarity between item pairs, to both spatial and temporal similarity, as well as semantic and orthographic features of the word stimuli. To manage the number of competing models, we did not pursue a full factorial approach to modelling intrusions. Instead, we build upon the Popov et al. (2021) finding of a temporal effect and introduce additional components in sequence to evaluate the improvement in model fit with each addition.

A secondary aim was to investigate if the simultaneous presentation of information (i.e. the presentation of a word in the location) or the sequential presentation of source and item information (i.e presenting the source location, followed by the item as in Harlow & Donaldson, 2013; Zhou et al., 2021) affected source responding. To this end, we experimentally manipulated the format of source presentation across participants.

In Experiment 1, we found qualitative improvements in model fit by introducing successively sophisticated models of intrusions between items, ranging from a pure guessing model with no intrusions to model with a spatiotemporal similarity gradient determining intrusion probability. However, quantitative evidence for the more sophisticated models 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 further elaborations including semantics and orthography in determining similarity between items.

# Experiment 1

## Method

### Participants

In Experiment 1, 10 participants were recruited online through the University of Melbourne undergraduate research experience program and 40 participants were recruited via *Prolific*, an online participant recruitment platform. 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 consisted of words generated from the SUBTLEXus database, filtered for words with a length of four letters, and with frequency ratings between 1 and 300, which represents the number of times the word appears in the corpus of 51 million words. Words were displayed in size 24 point “Courier New” white font positioned in the center of a uniform mean luminance field. The choice of a monospaced font and the restriction of words to strictly four letters were to ensure stimuli always occupied a consistent amount of space on the screen. Software written in Javascript using jsPsych (deLeeuw, 2015) controlled stimulus presentation and recorded responses.

### Procedure

Participants completed the experimental tasks over three sessions. Each of the three sessions consisted of 120 trials, which was broken up into 12 blocks of ten items each. Blocks consisted of a study phase, a math 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. There were two conditions in this experiment, a simultaneous study condition and a sequential study condition, with all other phases being identical between the conditions. Participants were randomly allocated to either the simultaneous or the sequential presentation condition when beginning session one of the experiment, which would be the same for all subsequent sessions for that participant.

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

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

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

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

Finally, in the source memory retrieval task, participants were cued with the words for 1500 ms, and then indicated the recalled location by a moving the mouse from the starting point in the centre of the circle to a point on the circumference of the response circle. There was no time limit on the decision task. A schematic for one trial in each of the phases is shown in Figure 2.

Figure 2

*Schematic of display presented to the participant in one trial in each phase of the experiment.*

A picture containing clock

Description automatically generated

## Results

In this section, we step through incremental elaborations of a model of response errors, and then repeat these steps with the circular diffusion model that models response error and response time data. First, we assess whether there is a difference between the sequential and simultaneous presentations of source and item, in terms of both recognition and source judgements. Second, we consider the contribution of intrusions and compare the predictions of a pure guessing model with a pure intrusion model, as well as a hybrid model with intrusions and guessing. Third, we introduce a more sophisticated intrusion component to the model that is sensitive to the temporal or spatiotemporal similarity between items when determining pairwise intrusion probabilities. Finally, we repeat these steps with the circular diffusion model to evaluate diffusion analogs of the competing models in a richer data space.

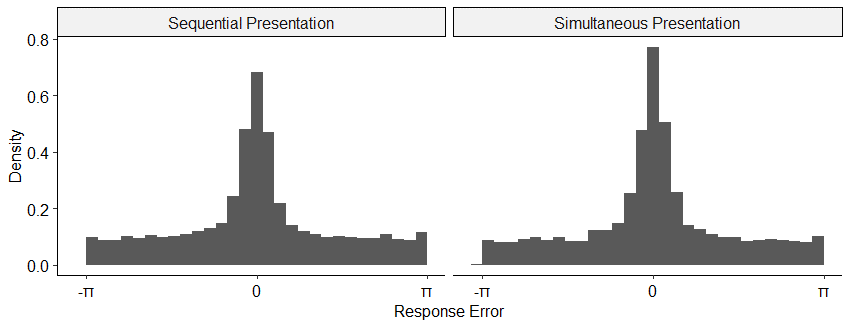
### Data Exclusion

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

With regard to performance in the source judgments, response error averaged within and compared between the simultaneous (M = .009, SD = 1.37) and sequential (M = .002, SD = 1.43) groups were not significantly different *t*(12460) = .28, *p* = .773. This can be confirmed visually by comparing the distributions of response error in the two conditions (Figure 3)

Figure 3   
*Normalized Histograms of Source Error in Sequential and Simultaneous Presentation Conditions*



Subsequent modelling analyses were conducted on an individual level, and significance tests on resultant parameter estimates between the presentation conditions were also not significant. These analyses are provided as supplementary material and commentary on the modelling will not make further reference to the presentation manipulation.

### Evidence of Intrusions

While guesses and intrusions will both appear uniform relative to the target on each trial, the two can be distinguished by examining the distance between responses and each of the non-target items on each trial (Bays et al., 2009). 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 4. We will subsequently refer to this analysis as *recentering* the data, as it is equivalent to recentering response errors relative to the non-target items.

Figure 4

*Response Angles Relative to Non-targets*

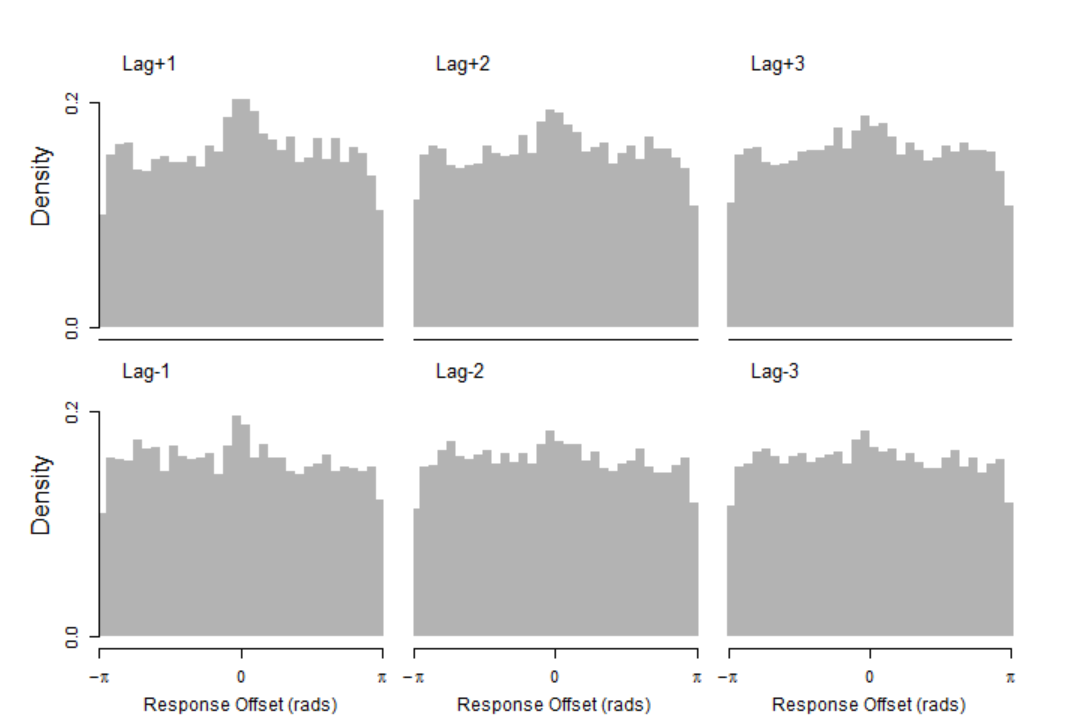
Chart, histogram

Description automatically generated

Figure 5 splits the recentered data by the lag and direction of the intrusion for each trial. Central tendency, and hence evidence for intrusions, is stronger in the forwards direction and decreases with higher absolute lag.

Figure 5

*Response Angles Relative to Non-targets, Split across Lag and Direction*



### Response Error Models

Our modelling strategy was to start with a two-component mixture model equivalent to Zhang and Luck (2008), and then introduce successive elaborations on the intrusion component to make it sensitive to similarity first in terms of temporal, then spatial similarity of presentation, and finally semantic and orthographic features of the stimuli. The same stepwise process was then taken with the circular diffusion model, using the same calculations to weight intrusion probability by the various kinds of similarity, using the Zhou et al. (2021) two-component circular diffusion model as a core in place of the Zhang and Luck (2008) model. The models are formally described in the sections to follow. In addition, we implemented variations of some models with allowances such as different parameters for primacy and recency items, and additive and multiplicative combinations of similarity when calculating intrusion probabilities. For ease of presentation, we have excluded these variants in this text, but code for all the models is available at [link here] and are provided as supplementary material.

### Model 1: Pure Guessing

As previously described, the Zhang and Luck (2008) model expresses the idea that responses are generated from a mixture of two process, one which is target-driven with Gaussian error, and another which is driven by guesses made at random:

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

where in the target-driven component, represents the target angle, is the reported angle, and represents a Von Mises distribution with a mean of 0 and a standard deviation of . The probability that a response is a guess is represented by .

### Model 2: Pure Intrusions

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

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

where the probability of an intrusion occurring is represented by , and the angle associated with the *i*th intruding item is represented by . Note that of the *m* non-target items, the probability of a particular non-target intruding is equal. We report fits of a model which allows different values of *δ* for target and intrusion von Mises distributions.

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

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

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

In contrast to Models 4 and 5 where the probability of a given non-target item intruding is dependent on its similarity to the target, intrusions in Model 3 all occur with equal probability. We refer to this feature of the model as a flat intrusion gradient.

Models were fit on an individual level data, and the relative performance of the models summed over participants is shown in Table 1. Performance is evaluated on the basis of AIC weights, which is interpretable as the probability that a given model is correct for that participant (Burnham & Anderson, 2002). For the majority of participants, Model 3 (Intrusion + Guess) is heavily preferred over simpler models with guessing or intrusion components exclusively, and more sophisticated models with contextual gradients on the probability of each intruding item. Table 1 shows AIC values summed over participants, in which Model 3 is also preferred. Models 4 and 5, which incorporate temporal and spatial similarity gradients on intrusion probabilities, are introduced in the following sections.

Table 1

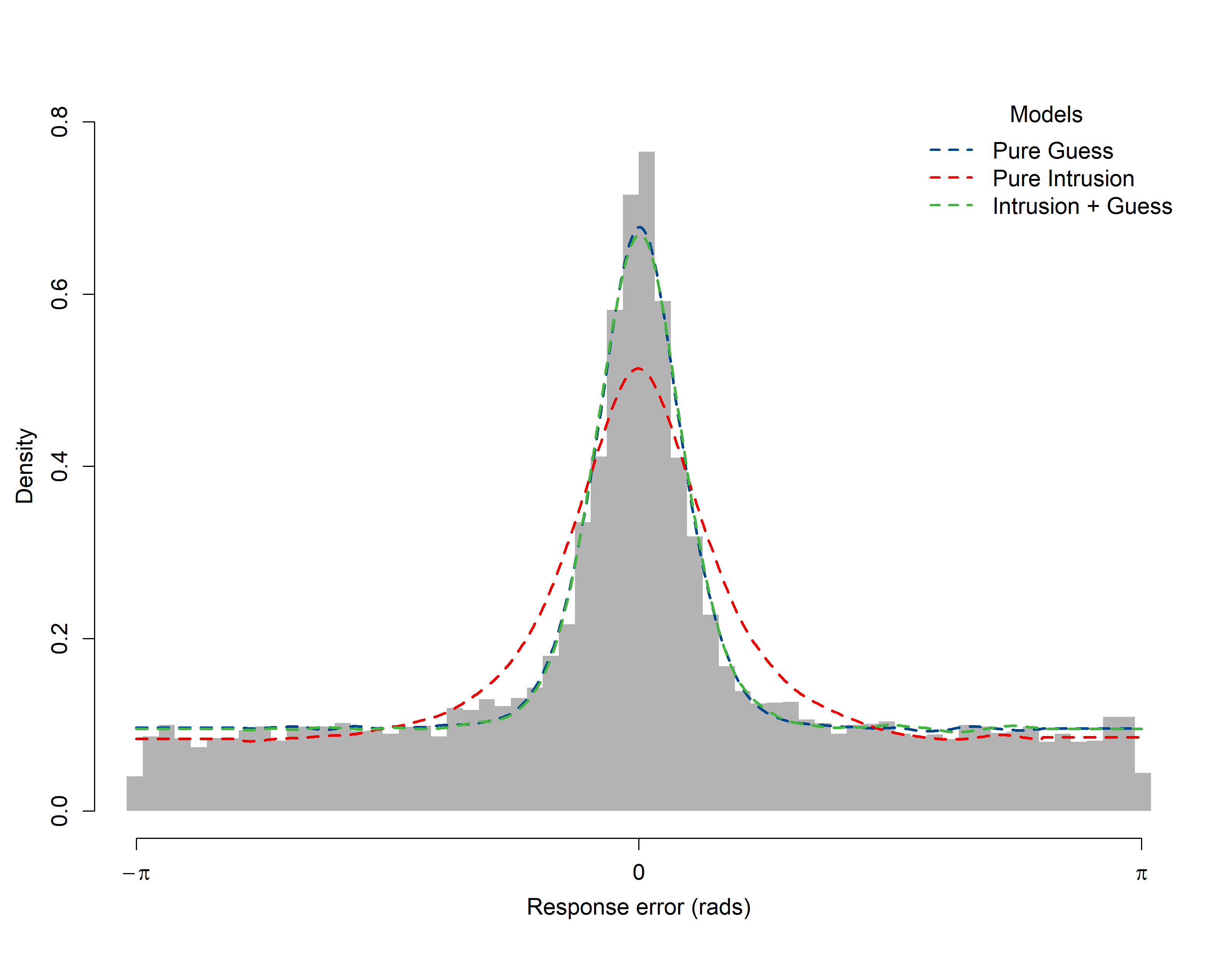
*AIC Values Summed Over Participants*

|  |  |  |  |
| --- | --- | --- | --- |
| Model Name | Parameters | ΣAIC | ΔΣAIC |
| 1. Pure Guess | 2 | 37338.77 | 276.86 |
| 1. Pure Intrusion | 3 | 38178.07 | 1116.16 |
| 1. Intrusion + Guess (Flat) | **4** | **37061.91** | **0** |
| 1. Temporal Gradient | 7 | 37176.82 | 114.91 |
| 1. Spatiotemporal Gradient | 9 | 37237.68 | 175.77 |

The reason for underperformance of the Pure Intrusion model compared to the Pure Guess and Intrusion + Guess models can be seen in Table 2, which shows the average estimated parameter for each model. The lower value of precision for the memory component *δ1* results in the Pure Guess model underestimating the peak of response error distribution around 0 seen in Figure 6. This suggests that a substantial portion of errors are not associated with intrusions, and that an additional source of error like the uniform guessing component of the other models is required to simultaneously account for these errors and the precision of target responses.

Figure 6

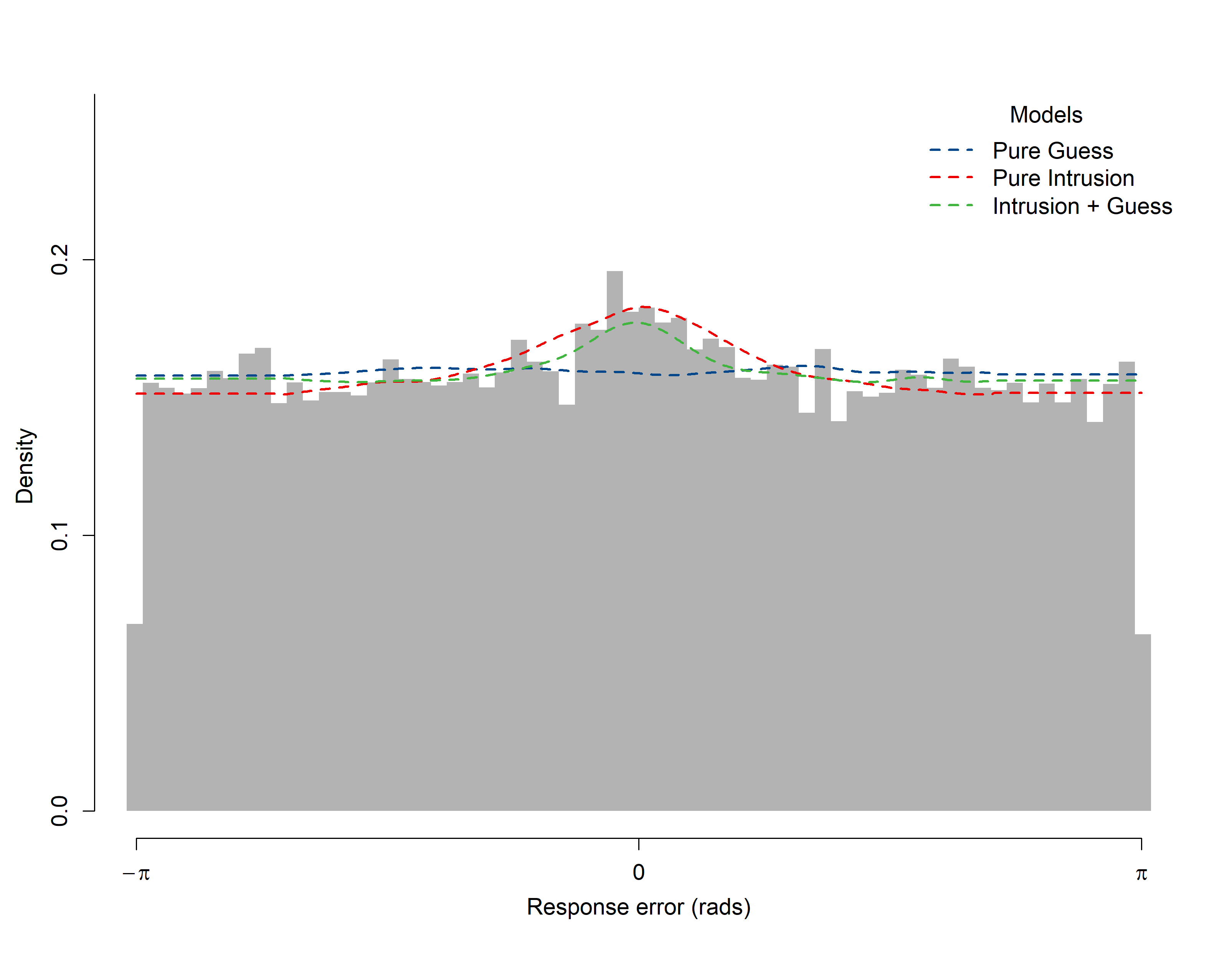
*Comparison of Pure Guess and Pure Intrusion Models to Response Error data*



However, as shown in Figure 7, the Pure Guess model does not predict the central tendency seen in the response error recentered on non-targets. The Intrusions + Guess model, with both guessing and intrusion components, is able to produce both patterns of data at the same time.

Figure 7

*Model Fits to Distances between Response Angles and Non-Target Angles*



### Model 4: Temporal Similarity Gradient

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

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

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

The probability of an intrusion occurring on a trial is the sum of temporal similarity values over all the available non-target lags for the study list position of the target. Because the possible lags are different for each position in the study list, the summed probability of intrusions also varies across trials. In the models we present, 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 trials:

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

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, which was not found to improve the fit of the model. We reflect on the plausibility of these assumptions and the limitation of mixture modelling this ambiguity reflects in the discussion section.

### Model 5: Spatiotemporal Similarity Gradient

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

|  |  |
| --- | --- |
|  | (7) |
|  | (8) |

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 similarity between the target and intruding angles *l*:

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

as with temporal similarity, we assume that spatial similarity decreases exponentially with distance, which in this case is the circular distance between the two angles. The relative contribution of temporal and spatial similarity in determining the probability of a particular non-target item intruding is weighted by . Naturally, intrusion responses from near non-targets will be associated with lower error relative to the target than intrusions from far non-targets. Therefore, as increases in Model 5, overall response error decreases. The best fitting parameter values for each model, estimated separately for each participants and then averaged across participants, is shown in Table 2.

Table 2

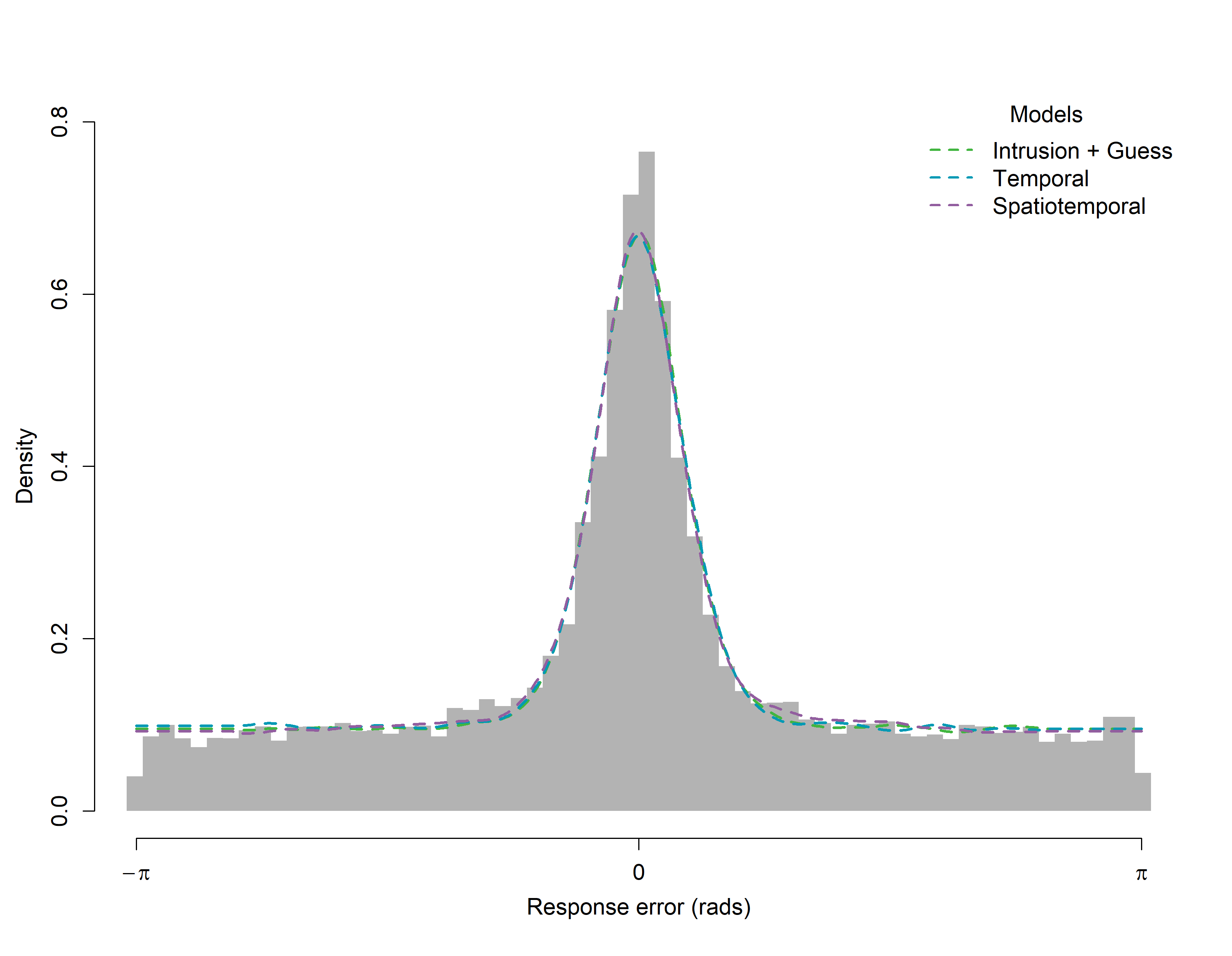
*Average Parameter Value Estimates for Each Model*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | *δ1* | *δ2* | *β* | *γ* | *κ* | *λ1* | *λ2* | *ζ* | *ρ* |
| 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 |

The Pure Guess and Intrusion + Guess models agree on the proportion of non-target responses (*β* = 0.60 in Model 1, *β + γ ≈* 0.60 in Model 3), but contrary to our expectations, the inclusion of temporal and spatiotemporal gradients in the intrusion component (in Models 4 and 5 respectively) did not further decrease the proportion of guesses relative to the flat gradient in Model 3.

Figure 8

*Comparison of Intrusion + Guess, Temporal, and Spatiotemporal Gradient Model Fits.*

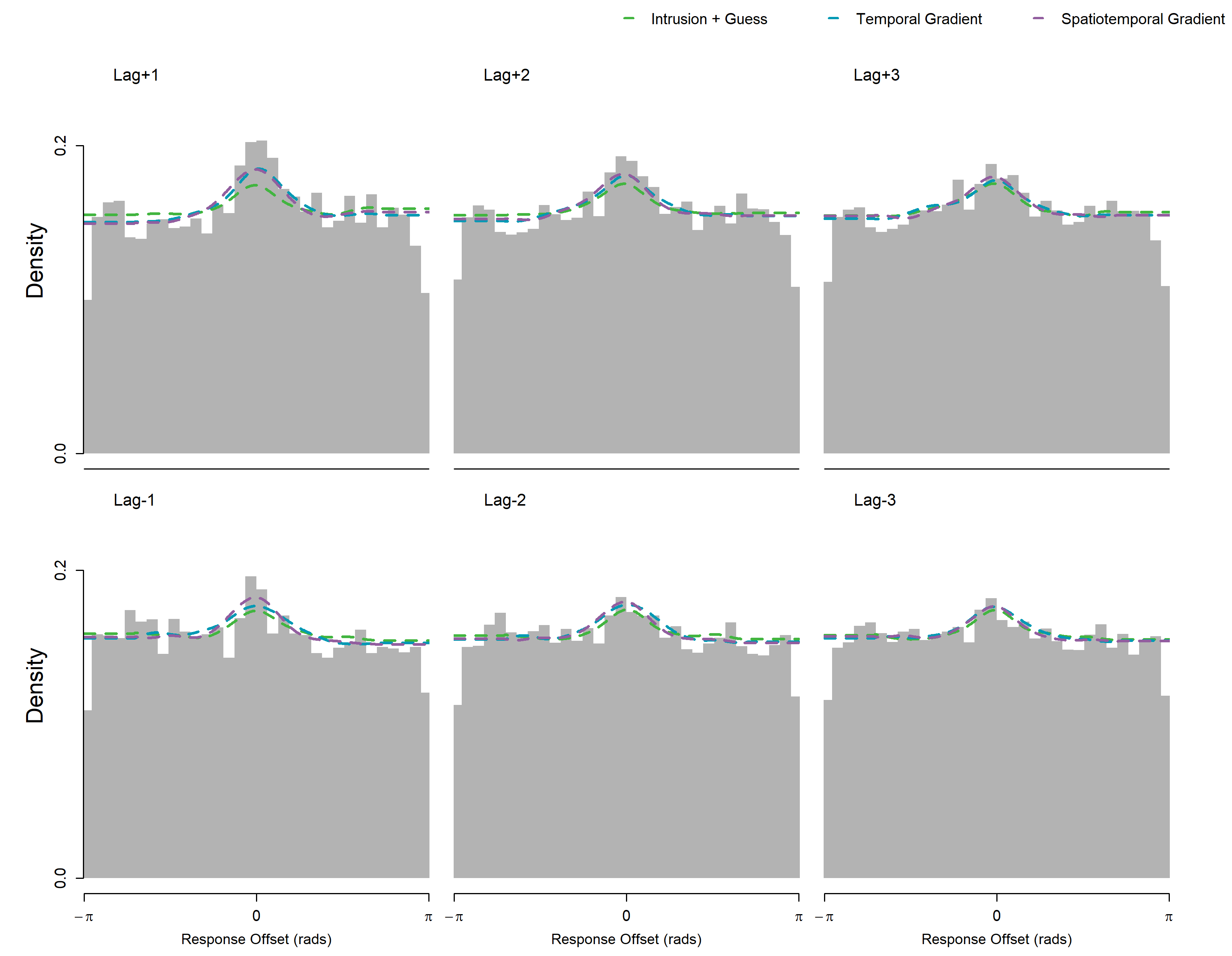


*Note*. All three models have both an intrusion and guessing component. In the Intrusion + Guess model, all non-targets are equally likely to intrude, while in the temporal and spatiotemporal models, intrusions are individually determined by the respective similarity gradients.

Models 3, 4, and 5 make visually indistinguishable predictions about the overall distribution of response errors (Figure 8). Instead, the effect of different intrusion probability gradients can be seen in the recentered data in Figure 9. Because Model 3 assumes that intrusions are equally likely from all non-target items, there is no relationship between lag magnitude or direction and how pronounced the central tendency is in the recentered data. In contrast, Models 4 and 5 predict fewer intrusions from greater lags and from backwards lags, a pattern which is present in the data (Figure 9).

Figure 9

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



Another qualitative advantage of the gradient models (Models 4 and 5) is that they predict a serial position effect, with lower response error for items at the start and end of the study list (Figure 10). The reason the gradient models make this prediction is the effect of the boundary at the ends of the list. For example, given that the greatest proportion of intrusions come from a lag of +1, then naturally the summed probability of intrusions is lowest for trials in which no items appear immediately after the target, i.e. the final trial in position 10.

Figure 10

*Average Response Error Across Target Serial Positions*

Chart, line chart, scatter chart

Description automatically generated

*Note*. Model predictions are represented by a loess curve through the average error of simulated data at each serial position.

### Circular Diffusion Models

To assess the models in a richer data space, we implemented diffusion analogs of each of response error models. The parameterization of the full intrusion diffusion model is as follows: mean drift is are represented by *μ*, which is normally distributed with standard deviation *η,* which reflects across-trial variability in evidence quality. We assume that memory strength differs between target and non-target responses, and so these parameters were estimated separately for the memory component (*μ1*, *η1*) and the intrusion component (*μ2*, *η2*), however, the two components share a single decision criterion (*a1*) because we assume that the decision process is blind to the identity of the item driving it. Slow errors are typically found as the result of a mixture of high and low drift trials due to drift rate variability. In the intrusion models, not only do drift rates vary across trials, but intrusion responses are assumed to be associated with lower mean drift rates than target responses, and so the prediction of a slow error effect is a consequence of how intrusions are represented in the models. The uniform guessing component was implemented as a third diffusion process with a mean drift of 0 and a separate decision criterion (*a2*), reflecting a state in which no information is driving the decision process, which requires less total evidence to generate a response than information-driven trials. Finally, non-decision time (*Ter*) is added to response times to represent the assumption that RTs are the sum of the duration of the decision process as well as other processes, such as encoding and the response itself. For a more detailed description of the circular diffusion model, see Smith (2016). The parameters governing the mixture of memory, guess, and intrusion components are the same as in the response error models previously described. The parameterization of the diffusion models, as well as the AIC values summed over all participants, are summarized in Table 3.

Table 3

*Diffusion Model Parameterization*

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Parameters (Number)** | **ΣAIC** | **ΔΣAIC** |
| 1. Pure Guess | *μ1, η1, a1, a2, Ter,β* **(6)** | 47611.67 | 1819.00 |
| 2. Pure Intrusion | *μ1, η1, μ2*, *η2, a1, Ter,γ* **(7)** | 46512.06 | 719.39 |
| 3. Intrusion + Guess (Flat) | *μ1, η1, μ2*, *η2, a1, a2, Ter,β, γ* **(9)** | 45850.07 | 57.41 |
| 4. Temporal Gradient | *μ1, η1, μ2*, *η2, a1, a2, Ter,β, γ, κ, λ1, λ2* **(12)** | 45988.75 | 196.09 |
| 5. Spatiotemporal Gradient | *μ1, η1, μ2*, *η2, a1, a2, Ter,β, γ, κ, λ1, λ2, ζ, ρ* **(14)** | 45792.67 | 0.00 |

On the basis of AIC values summed over participants, the spatiotemporal diffusion model is preferred over the other diffusion model variants. The average estimated values of each parameter are shown in Table 4.

Table 4

*Diffusion Model Parameter Estimates*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | *μ1* | *μ2* | *η1* | *η2* | *a1* | *a2* | *γ* | *β* | *κ* | *λ1* | *λ2* | *ζ* | *ρ* | *Ter* |
| 1 | 1.78 |  | 0.21 |  | 2.78 | 1.68 |  | 0.62 |  |  |  |  |  | 0.18 |
| 2 | 2.41 | 2.14 | 0.26 | 0.58 | 1.86 |  | 0.44 |  |  |  |  |  |  | 0.16 |
| 3 | 3.32 | 2.70 | 0.30 | 0.32 | 2.10 | 1.29 | 0.27 | 0.28 |  |  |  |  |  | 0.19 |
| 4 | 3.25 | 1.78 | 0.24 | 0.27 | 2.07 | 1.21 | 0.31 | 0.38 | 0.49 | 0.80 | 0.94 |  |  | 0.19 |
| 5 | 3.51 | 2.32 | 0.19 | 0.29 | 2.17 | 1.26 | 0.16 | 0.35 | 0.64 | 0.87 | 0.76 | 0.39 | 0.56 | 0.19 |

Both the distribution of response errors, Figure 11, and the joint distribution of error and RT in Figure 12 demonstrate that the two-component models (Models 1 and 2) overestimate response error relative to the three-component models (Models 3, 4, and 5).

Figure 11

*Diffusion Model Fits to Response Error and Latency*

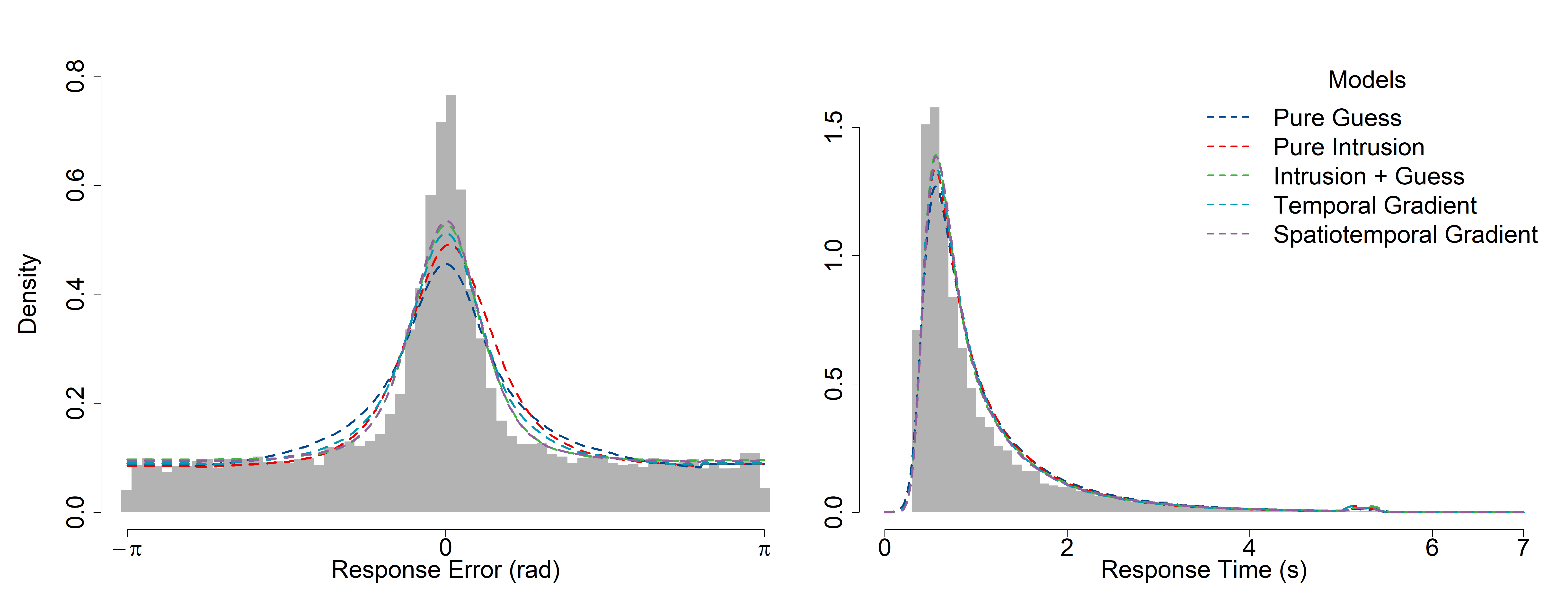
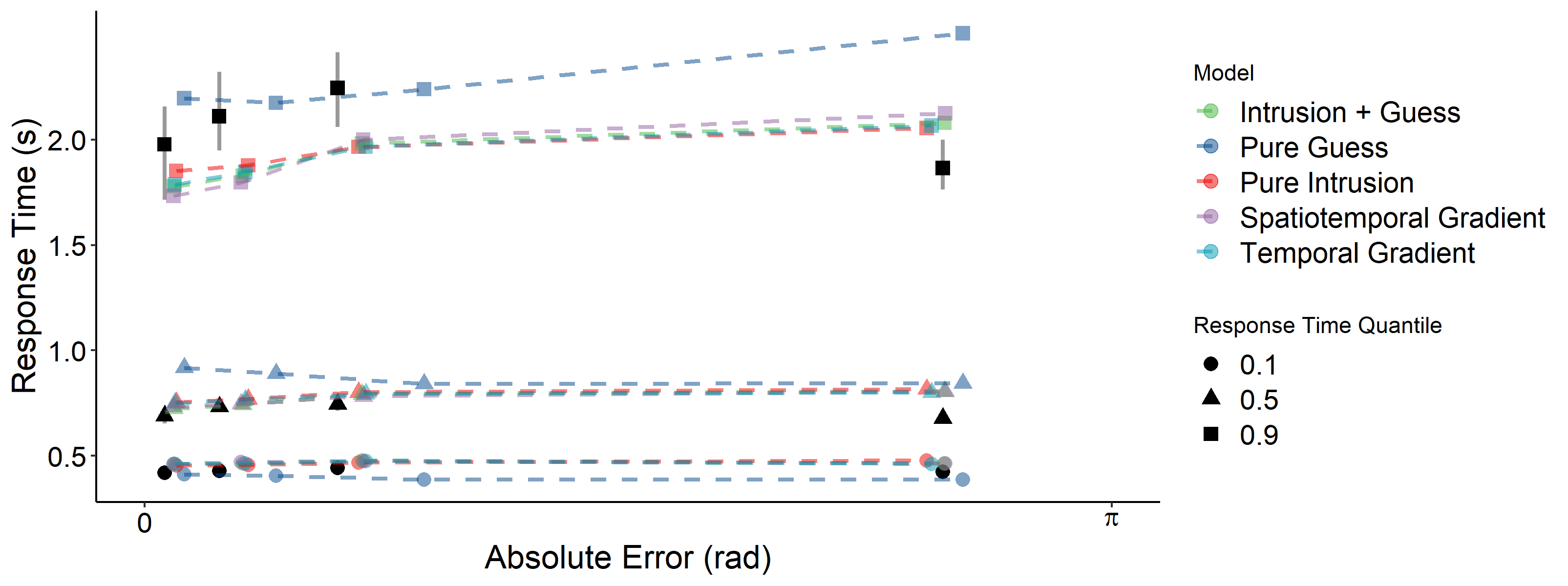


Figure 12

*Model Fits to Joint Response Error and Time Quantiles*



*Note*. Grey lines represent the 95% confidence interval around the observed response time quantiles. The error quantiles, which are unlabeled in the figure, are the .1, .3, .5, and .9 quantiles moving from left to right along the x-axis. The error quantiles are the upper bound defining the bin of responses for the corresponding response time quantiles, which are stacked vertically, while the lower bound for each bin is the next-lowest error quantile. There is no lower bound for the 0.1 error quantile bin.

## Discussion

In Experiment 1, we found unambiguous support for a model with both intrusions and guessing components over models with only one of these error components. The sensitivity of the intrusion component to similarity between target and intruding items was less clear. We found successive qualitative improvement when intrusion probabilities were determined by temporal and spatiotemporal similarity gradients, compared to the base three-component Intrusion + Guess model in which all non-target items are equally likely to intrude. When fit to response error data alone, the overall likelihood of the three-component models were close enough that marginal improvements with the more sophisticated gradient models were outweighed by the parameter penalty incurred by the similarity gradients.

When diffusion analogs of each model were additionally constrained by also fitting RT data, the model predictions were differentiated, resulting in a quantitative advantage for the spatiotemporal gradient. One explanation for the mixed results, both in terms of qualitative and quantitative response error evidence, as well as response error and joint data, is that there were simply insufficient observations in the participant-level data to support 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.

# Experiment 2

## Method

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

### Participants

In Experiment 2, participants were recruited solely via Prolific. Of the 10 participants recruited, four participants did not finish all sessions of the experiment, and one participant was excluded as the Rayleigh test indicated no deviance from uniform responding, leaving a final sample of five participants included for the analyses.

### Procedure

Source and item information was only presented simultaneously for all participants in Experiment 2. The number of sessions each participant completed was increased from three sessions to 10 sessions.

## Results

In these modelling analyses, we excluded the two-component Pure Intrusion and Pure Guess models as results from Experiment 1 showed that a model with both intrusions and guessing is necessary to produce the pattern of response errors and recentered errors in the source memory task. In addition, we introduce models that incorporate orthographic and semantic similarity between targets and non-targets to determine intrusion probabilities.

### Model 6: Orthographic Model

In the orthographic model, orthographic similarity between the target and a non-target word is represented by *o* and is calculated from the Levenshtein distance of the two four-letter strings, and then weighted against the spatiotemporal similarity of the presentation context given in (8). The resultant weight then determines the individual probability of a given non-target item intruding, as in the previous models expressed in (7):

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

In words, the probability of a non-target item intruding is a weighted product of the spatiotemporal similarity of the two items at presentation (which is itself a weighted product of temporal and spatial similarity), and the orthographic similarity between the non-target and target word.

### Model 7: Semantic Model

The semantic model substitutes orthographic similarity in Model 6 for semantic similarity between target and non-target words. Semantic similarity, *s*, is defined as the cosine similarity between vector representations of each word derived from *fast-Text* models which were pre-trained on multiple corpora of natural text (Mikolov et al., 2017):

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

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

In the four-factor model, both semantic and orthographic components are combined multiplicatively with an additional parameter, *ψ,* governing the weight of semantic similarity relative to orthographic similarity:

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

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

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

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

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

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

Table 5

*Response Error Model Parameterization*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Parameter** | | **Description** | | |
| *δ1* | | Precision, memory | | |
| *δ2* | | Precision, intrusion | | |
| *β* | | Proportion of uniform guesses | | |
| *γ* | | Proportion of intrusion responses | | |
| *κ* | | Temporal gradient asymmetry | | |
| *λ1* | | Temporal similarity decay, forwards | | |
| *λ2* | | Temporal similarity decay, backwards | | |
| *ζ* | | Spatial similarity decay | | |
| *ρ* | | Spatial vs. Temporal similarity weight | | |
| *χ* | | Spatiotemporal vs. Semantic/Orthographic weight | | |
| *ψ* | | Semantic vs. Orthographic weight | | |
| **Model** | **Parameters** | | **Number of Parameters** | |
| 3. Intrusion + Guess (Flat) | *δ1, δ2, β,γ* | | | 4 |
| 4. Temporal Gradient | *δ1, δ2, β,γ, κ, λ1, λ2* | | | 7 |
| 5. Spatiotemporal Gradient | *δ1, δ2, β,γ, κ, λ1, λ2, ζ, ρ* | | | 9 |
| 6. Orthographic Gradient | *δ1, δ2, β,γ, κ, λ1, λ2, ζ, ρ, χ* | | | 10 |
| 7. Semantic Gradient | *δ1, δ2, β,γ, κ, λ1, λ2, ζ, ρ, χ* | | | 10 |
| 8. Four Factor (Additive) | *δ1, δ2, β,γ, κ, λ1, λ2, ζ, ρ, χ, ψ* | | | 11 |
| 9. Four Factor (Multiplicative) | *δ1, δ2, β,γ, κ, λ1, λ2, ζ, ρ, χ, ψ* | | | 11 |

The quantitative fit of each of these models are presented in Table 6. To show the relative strength of evidence for each model at an individual and group level, AIC weights are presented alongside AIC values, which are interpretable as conditional probabilities for each model (Wagenmakers & Farrell, 2004).

Table 6

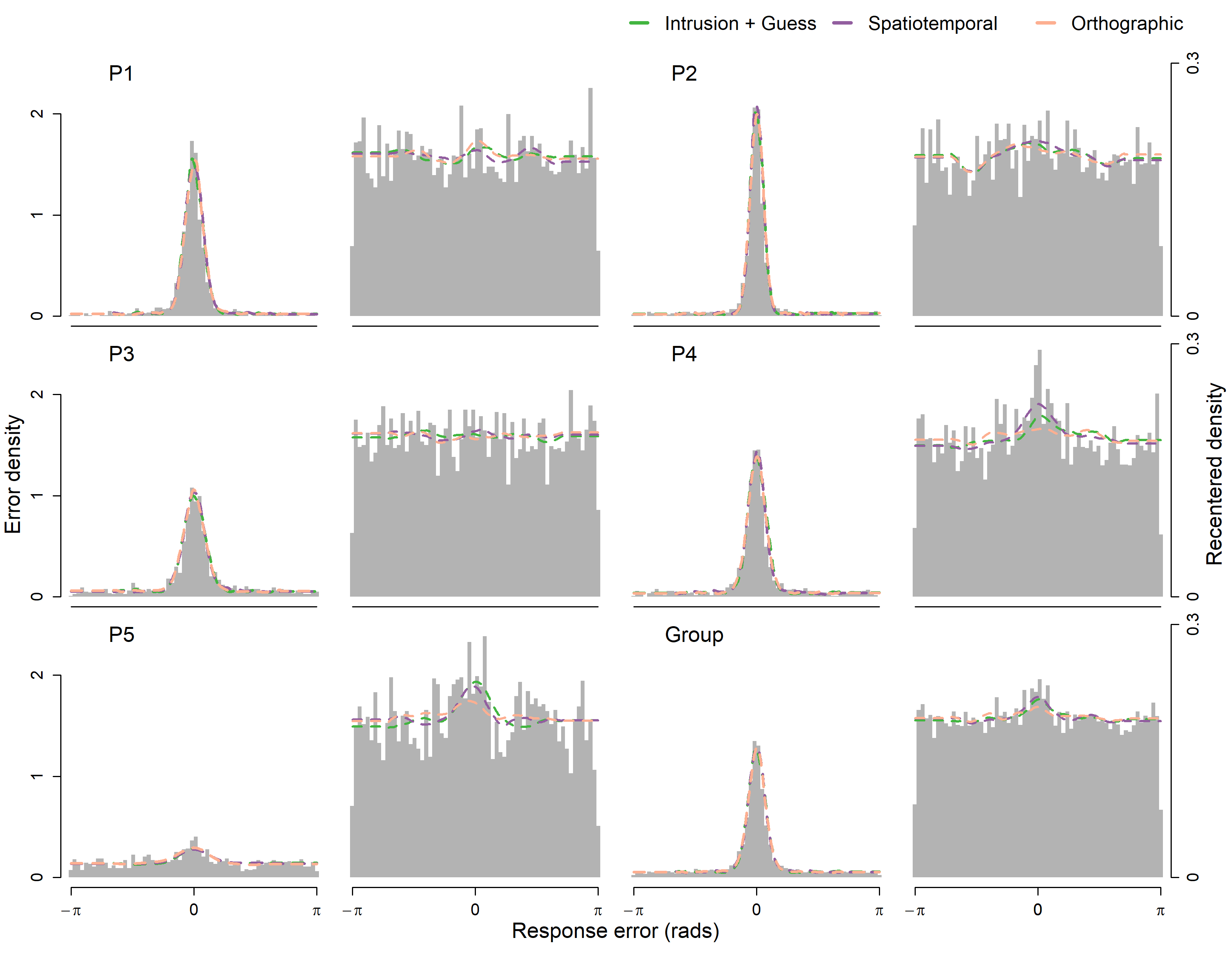
*AIC Weights for Individual and Group-level Response Error Model Fits*

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

At an individual level, the spatiotemporal model 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, the temporal and orthographic model are more competitive. The orthographic model is preferred outright for Participant 2, while for Participant 5 the simplest three-component model with uniform intrusion probabilities is preferred. The models in which intrusion probabilities are affected by semantic similarity (7, 8, and 9) were not well supported for any of the datasets. Figure 13 shows the graphical fits of the models to each participant-level dataset, excluding the four-factor models.

Figure 13

*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 each of the models in Figure 13 are difficult to distinguish. One concern in comparing models is the diagnosticity of model selection when the models make similar predictions. To evaluate the extent to which our models mimic each other, we conducted a model recovery exercise, which we limited to the spatiotemporal, orthographic, and semantic models. 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 generative 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. Because the effect of semantic similarity is minimal in this dataset, the estimated value of *χ* is so low that simulated data generated from the fitted parameters are not distinguishable from the spatiotemporal model (average parameter estimate values are presented in Table 7).

Table 7

*Parameter estimates for each model, averaged across participants*

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Parameter Average | | | | | | | | | | |
| *δ1* | *δ2* | *β* | *γ* | *κ* | *λ1* | *λ2* | *ζ* | *ρ* | *χ* | *ψ* |
| 3 | 22.14 | 12.49 | 0.21 | 0.14 |  |  |  |  |  |  |  |
| 4 | 20.83 | 11.12 | 0.22 | 0.35 | 0.59 | 0.66 | 0.36 |  |  |  |  |
| 5 | 22.93 | 10.15 | 0.20 | 0.08 | 0.79 | 2.03 | 1.11 | 0.58 | 0.80 |  |  |
| 6 | 23.31 | 11.05 | 0.19 | 0.16 | 0.74 | 2.07 | 0.49 | 0.81 | 0.19 | 0.23 |  |
| 7 | 23.40 | 10.97 | 0.20 | 0.12 | 0.66 | 1.55 | 0.80 | 0.81 | 0.28 | 0.10 |  |
| 8 | 22.19 | 10.37 | 0.20 | 0.04 | 0.75 | 2.30 | 2.34 | 0.83 | 0.32 | 0.37 | 0.20 |
| 9 | 21.93 | 15.88 | 0.30 | 0.20 | 0.75 | 1.89 | 1.55 | 0.88 | 0.34 | 0.50 | 0.01 |

### Diffusion Models

As with Experiment 1, we also compared diffusion versions of each model. Table 8 shows the AIC and AIC weights for the individual-level diffusion fits. The spatiotemporal diffusion model is preferred for all five participants, although the difference in quality of fit is smaller between the flat gradient model and the spatiotemporal gradient model for Participant 1 than for the other participants.

Table 8

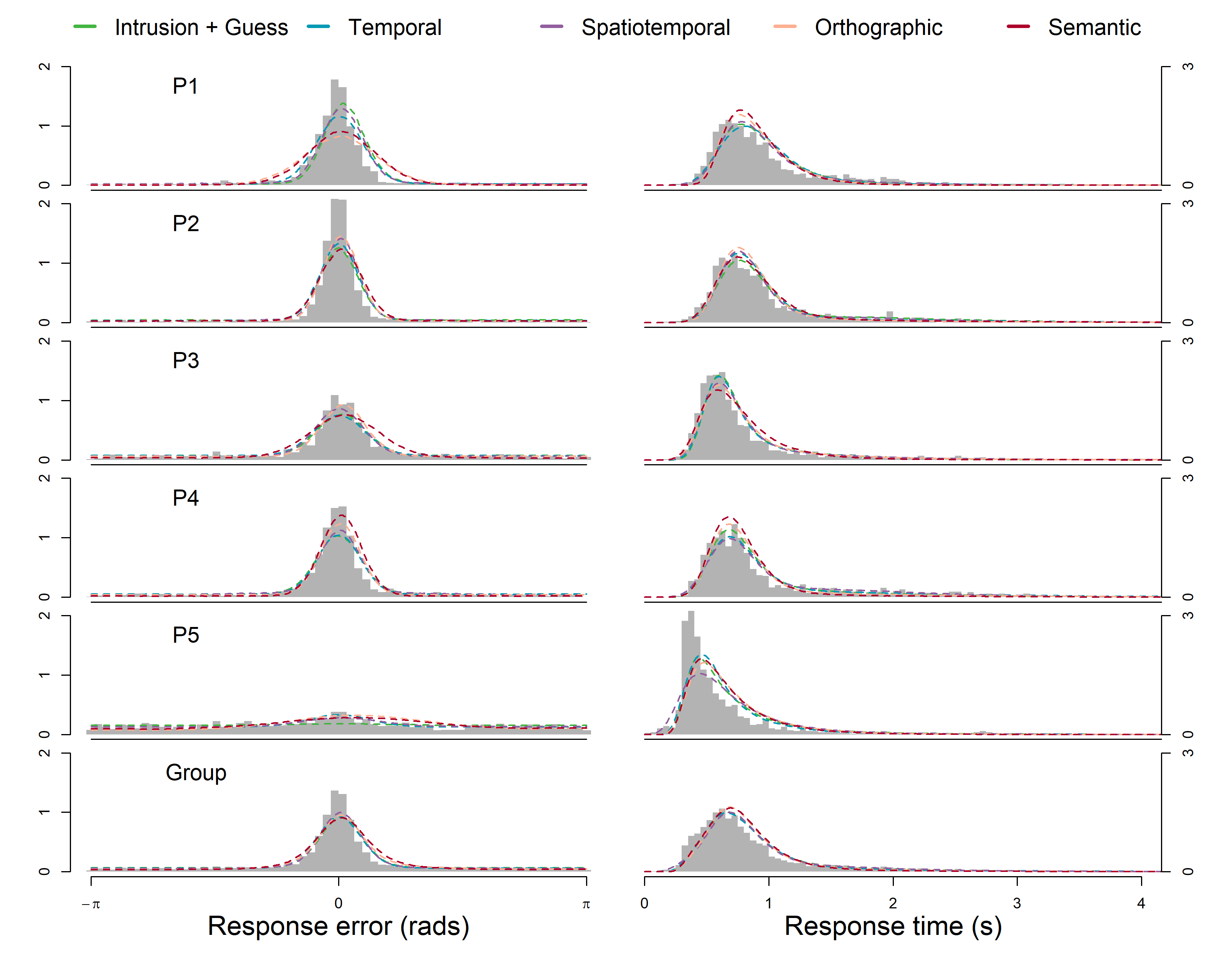
*Experiment 2 Diffusion Model AIC Comparison*

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

The graphical fit of the models to response error and times is shown in Figure 14. The additive and multiplicative four-factor models (models 8 and 9) performed worse than the three-factor semantic and orthographic models and are excluded from Figure 14 to better present the other model predictions.

Figure 14

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



The relationship between response time over response error is most clearly demonstrated at a group level in Figure 15, which plots response time quantiles for the data binned by response error quantiles. The semantic and orthographic models underestimate the number of responses made with high error. The average estimate for the proportion of guesses *β* is lower for these models (Table 9), resulting in a 0.9 error quantile prediction lower than the observed data or the models without orthographic or semantic intrusion factors. Conversely, the flat and temporal models overestimate high error responses, attributable to high estimates of *β*. The spatiotemporal model makes the closest prediction in terms of high error responses. Notably, all models under consideration misfit the .9 RT quantiles for the three most accurate error bins. This can be interpreted as a proportion of accurate responses which are slower than the models predict. In particular, the flat, temporal, and spatiotemporal gradient models overpredict the magnitude of the slow error effect, while the orthographic and semantic models uniformly underpredict the slowest RTs across the entire range of errors. There are substantial changes in RT across sessions, such that RTs in the first sessions tend to be slower than later sessions for most participants, which may explain why we do not observe a miss of this magnitude in Experiment 1, which had fewer subsequent sessions.

Figure

*Model fits to group-level response error and time quantiles*

Chart, scatter chart

Description automatically generated

Table 9

*Average Parameter Estimates*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Average Parameter Estimates | | | | | | | |
| *μ1* | *μ2* | *η1* | *η2* | *a1* | *a2* | *γ* | *β* |
| 3 | 3.95 | 1.51 | 0.32 | 0.39 | 2.83 | 1.36 | 0.12 | 0.30 |
| 4 | 4.34 | 0.97 | 0.04 | 0.10 | 2.78 | 1.44 | 0.14 | 0.32 |
| 5 | 4.71 | 1.46 | 0.24 | 0.08 | 3.03 | 1.34 | 0.07 | 0.27 |
| 6 | 3.71 | 1.76 | 0.44 | 0.19 | 2.60 | 1.37 | 0.17 | 0.20 |
| 7 | 3.18 | 1.04 | 0.41 | 0.15 | 2.30 | 1.21 | 0.15 | 0.15 |
| 8 | 3.69 | 1.72 | 0.27 | 0.24 | 2.70 | 1.49 | 0.12 | 0.15 |
| 9 | 4.26 | 0.16 | 0.11 | 0.01 | 2.72 | 1.29 | 0.09 | 0.34 |
|  | *κ* | *λ1* | *λ2* | *ζ* | *ρ* | *χ* | *ψ* | *Ter* |
| 3 |  |  |  |  |  |  |  | 0.10 |
| 4 | 0.49 | 1.23 | 0.53 |  |  |  |  | 0.07 |
| 5 | 0.72 | 0.17 | 0.62 | 0.78 | 0.86 |  |  | 0.07 |
| 6 | 0.67 | 0.77 | 1.12 | 0.36 | 0.51 | 0.23 | 0.00 | 0.11 |
| 7 | 0.70 | 0.62 | 0.89 | 0.29 | 0.22 | 0.33 | 1.00 | 0.13 |
| 8 | 0.62 | 0.45 | 0.94 | 0.37 | 0.48 | 0.18 | 0.38 | 0.08 |
| 9 | 0.74 | 0.70 | 1.03 | 0.10 | 0.45 | 0.46 | 0.10 | 0.09 |

## Discussion

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

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

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

# General Discussion

Our main theoretical goal in this study was to investigate the contribution of intrusions from non-target items in a continuous-outcome source memory task, while also empirically testing the effect of sequential and simultaneous presentation of item and source information in the task. The main contribution of our work in introducing the similarity-based intrusion probability gradient models is that it represents a novel attempt to systematically model similarity effects in memory in a continuous domain. Not only do we fit these models to response error, but we also account for distributions of response times, and demonstrate the value of the additional constraint such an approach affords.

Firstly, despite concern raised by Harlow and Donaldson (2013) that the simultaneous presentation of item and source would introduce a methodological confound due to unitization between the two, we found no differences between participants in the sequential or simultaneous presentation conditions in Experiment 1, which suggests that the binding of items to locations occur in both presentation modalities which motivated our decision to simultaneously present all items in their locations in Experiment 2.

## Intrusions

In Experiment 1, we found that intrusions accounted for some but not all error responses. While Bays et al. (2009) were able to eliminate the need for a uniform guessing component to account for errors in visual working memory, some proportion of high error responses in our source memory task appear not to be associated with non-targets from the same study list. We found that a three-component model with both intrusions from non-targets and guesses was strongly preferred over two-components model with either guessing or intrusions in isolation, suggesting that both of these processes contribute to error. Ultimately, findings reinforce the position that source memory retrieval is thresholded and that participants guess when memory strength is subtreshold, though given our finding that intrusions account for some of these errors, prior estimates of the rates of guessing are likely overestimates (Harlow & Donaldson, 2013; Zhou et al., 2021).

### Methodological Implications

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

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

### Theoretical Implications

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

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

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

## Dynamic Rates of Guessing and Limitations of Mixture Modeling

One potential limitation of the family of models explored in the present study is we assumed that changes in the summed probability of intrusions across trials did not affect the probability of guesses, which remained constant. It may not always be reasonable to expect that the proportion of guesses remains the same across serial positions. To test this assumption in a coarse way, we implemented versions of the model where the parameter governing the proportion of guesses was separately estimated for the first and last items in the study list, but we did not find that these models made consistently different predictions from the base family of models and have therefore omitted them. To take another example, consider the potential interaction between recognition and intrusion probability where items that are not recognized do not intrude. In a list where no items are recognized, we would intuit that all responses should be guesses. A more disciplined approach requires a formal process model of how memory, intrusion, and guesses compete under different scenarios. This underscores the fundamental ambiguity of mixture models with more than two mixture components. A clear solution that could be explored in future work would be to implement the models introduced in this study in a race architecture, such that the target and all non-target responses are modelled as separate evidence accumulation processes that compete to be retrieved in the manner discrete multi-alternative decisions have been modelled (Roe et al., 2001; Zandbelt et al., 2014), with an additional process representing guesses (Hawkins & Heathcote, 2021). However, to date, no race model of evidence accumulation in a continuous domain exists.

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Table X  *Individual-level Response Error Model Comparison* | | | | | |
| Participant | Model AIC Weight | | | | |
| 1  Pure Guess | 2  Pure Intrusion | 3  Intrusion + Guess | 4  Temporal | 5  Spatiotemporal |
| 1 | 0.12 | 0.00 | **0.88** | 0.00 | 0.00 |
| 2 | **0.62** | 0.08 | 0.23 | 0.05 | 0.02 |
| 3 | **0.54** | 0.00 | 0.33 | 0.12 | 0.02 |
| 4 | 0.07 | 0.00 | **0.88** | 0.03 | 0.02 |
| 5 | 0.00 | 0.00 | **0.96** | 0.03 | 0.01 |
| 6 | 0.00 | 0.00 | 0.00 | 0.38 | **0.62** |
| 7 | 0.01 | 0.00 | **0.86** | 0.12 | 0.02 |
| 8 | 0.00 | 0.00 | 0.24 | 0.09 | **0.66** |
| 9 | 0.04 | 0.00 | **0.49** | 0.29 | 0.18 |
| 10 | 0.00 | 0.00 | **0.94** | 0.03 | 0.03 |
| 11 | 0.01 | 0.00 | **0.68** | 0.15 | 0.15 |
| 12 | **0.44** | 0.00 | 0.27 | 0.27 | 0.02 |
| 13 | **0.68** | 0.00 | 0.25 | 0.06 | 0.01 |
| 14 | 0.00 | 0.01 | **0.99** | 0.00 | 0.00 |
| 15 | **0.40** | 0.00 | 0.15 | **0.40** | 0.05 |
| 16 | 0.01 | 0.00 | 0.10 | **0.73** | 0.16 |
| 17 | 0.06 | 0.00 | **0.74** | 0.16 | 0.04 |
| 18 | **0.81** | 0.00 | 0.18 | 0.01 | 0.00 |
| 19 | 0.00 | 0.00 | **0.49** | 0.02 | **0.49** |
| 20 | 0.00 | 0.00 | 0.06 | **0.47** | **0.47** |
| 21 | 0.00 | 0.00 | **0.70** | 0.26 | 0.04 |
| 22 | 0.00 | 0.00 | **0.42** | **0.42** | 0.16 |
| 23 | 0.00 | 0.00 | **0.71** | 0.26 | 0.04 |
| 24 | 0.00 | 0.00 | **0.60** | 0.37 | 0.03 |
| 25 | 0.13 | 0.00 | 0.21 | **0.58** | 0.08 |
| 26 | 0.03 | 0.00 | **0.88** | 0.07 | 0.02 |
| 27 | **0.72** | 0.00 | 0.27 | 0.01 | 0.00 |
| 28 | 0.01 | 0.00 | **0.93** | 0.05 | 0.02 |
| 29 | **0.46** | 0.01 | **0.46** | 0.06 | 0.01 |
| 30 | 0.00 | 0.01 | **0.98** | 0.00 | 0.01 |
| 31 | 0.00 | 0.00 | 0.02 | **0.49** | **0.49** |
| 32 | 0.08 | 0.00 | **0.61** | 0.22 | 0.08 |
| 33 | **0.63** | 0.00 | 0.23 | 0.09 | 0.05 |
| 34 | **0.59** | 0.00 | 0.36 | 0.05 | 0.00 |
| 35 | **0.47** | 0.00 | **0.47** | 0.04 | 0.02 |
| 36 | **0.59** | 0.00 | 0.36 | 0.05 | 0.01 |

*Note*. The AIC weight can be interpreted as the relatively likelihood of a given model for each participant

The estimated parameter values (experiment 2, response error model), averaged across participants, are shown in Table 7.

Table 10

*Average Parameter Estimates of Response Error Models*

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | *δ1* | *δ2* | *β* | *γ* | *κ* | *λ1* | *λ2* | *ζ* | *ρ* | *χ* | *ψ* |
| 3 | 22.14 | 12.49 | 0.21 | 0.14 |  |  |  |  |  |  |  |
| 4 | 20.83 | 11.12 | 0.22 | 0.35 | 0.59 | 0.66 | 0.36 |  |  |  |  |
| 5 | 22.93 | 10.15 | 0.20 | 0.08 | 0.79 | 2.03 | 1.11 | 0.58 | 0.80 |  |  |
| 6 | 22.10 | 15.84 | 0.30 | 0.18 | 0.64 | 1.83 | 1.61 | 0.86 | 0.62 | 0.53 |  |
| 7 | 22.74 | 9.39 | 0.19 | 0.07 | 0.86 | 2.22 | 2.19 | 0.54 | 0.51 | 0.78 |  |
| 8 | 22.69 | 10.15 | 0.20 | 0.20 | 0.73 | 1.51 | 0.75 | 0.75 | 0.53 | 0.19 | 0.13 |
| 9 | 21.93 | 15.60 | 0.30 | 0.23 | 0.71 | 1.16 | 1.23 | 0.82 | 0.36 | 0.53 | 0.02 |

1. The von Mises distribution is a circular analogue of the Gaussian distribution. [↑](#footnote-ref-1)
2. We refer to erroneous responses driven by non-target items as intrusions, describing how words from non-target word-location pairs are intruding on the cued pair. These within-list intrusions are not to be confused with extra-list intrusions, or *protrusion* errors, which we do not expect to contribute to errors in our paradigm. [↑](#footnote-ref-2)
3. The authors note that Healey (1974) observed a symmetric transposition gradient and Madigan (1971) observed asymmetry in the opposite direction, potentially due to differences in response modality. [↑](#footnote-ref-3)