Title
Distinguishing guesses from fuzzy memories: Further evidence for item limits in visual working memory

Journal
Attention, Perception, & Psychophysics, 85(5)

ISSN
1943-3921

Authors
Ngiam, William XQ
Foster, Joshua J
Adam, Kirsten CS
et al.

Publication Date
2023-07-01

DOI
10.3758/s13414-022-02631-y

Distinguishing guesses from fuzzy memories: Further evidence for item limits in visual working memory

William X. Q. Ngiam[1,2], Joshua J. Foster[1,2], Kirsten C. S. Adam[3], Edward Awh[1,2]

[1]Department of Psychology, University of Chicago
[2]Institute of Mind and Biology, University of Chicago
[3]Department of Psychology, University of California San Diego

Corresponding author: William X. Q. Ngiam, 940 E 57th Street Chicago Illinois, United States, wngiam@uchicago.edu

Conceptualization: J.J.F., K.C.S.A., and E.A.
Data curation: W.X.Q.N.
Formal analysis: W.X.Q.N., J.J.F., K.C.S.A., and E.A.
Funding acquisition: E.A.
Investigation: J.J.F. and K.C.S.A.
Methodology: W.X.Q.N., J.J.F., and K.C.S.A.
Project administration: W.X.Q.N.
Software: W.X.Q.N., J.J.F., and K.C.S.A.
Supervision: E.A.
Validation: W.X.Q.N.
Visualization: W.X.Q.N. and J.J.F.
Writing - original draft: W.X.Q.N., J.J.F., K.C.S.A., and E.A.
Writing - review & editing: W.X.Q.N., J.J.F., K.C.S.A., and E.A

Note: This manuscript is a preprint and has not yet been peer reviewed or any post-submission improvements or corrections. The Version of Record of this article is published in *Attention, Perception, & Psychophysics*, and is available online at https://doi.org/10.3758/s13414-022-02631-y.

Abstract

There is consistent debate over whether capacity in working memory (WM) is subject to an item limit, or whether an unlimited number of items can be held in this online memory system. The item limit hypothesis clearly predicts guessing responses when capacity is exceeded, and proponents of this view have highlighted evidence for guessing in visual working memory tasks (e.g, Adam et al., 2017; Zhang & Luck, 2008). Nevertheless, various models that deny item limits can explain the same empirical patterns by asserting extremely low fidelity representations that cannot be distinguished from guesses. To address this ambiguity, we employed a task for which guess responses elicited a qualitatively distinct pattern from low fidelity memories. Inspired by work from Rouder et al. (2014), we employed an orientation WM task that required subjects to recall the precise orientation of each of six memoranda presented one second earlier. The orientation stimuli were created by rotating the position of a “clock hand” inside a circular region that was demarcated by four colored quadrants. Critically, when observers guess with these stimuli, the distribution of responses is biased towards the center of these quadrants, creating a “banded” pattern that cannot be explained by a low precision memory. We confirmed the presence of this guessing pattern using formal model comparisons, and we show that the prevalence of this pattern matches observers’ own reports of when they thought they were guessing. Thus, these findings provide further evidence for guessing behaviors predicted by item limit models of WM capacity.

Introduction

Visual working memory (WM) enables the "online" retention of information for ongoing perception and cognition. Although there is broad agreement that WM capacity is subject to strong limits, there has been persistent debate regarding the nature of those limits. When tasked with remembering a single visual feature such as the orientation of a line or the color of a square, observers can reproduce the remembered feature value with very little error (Zhang & Luck, 2008). However, when the number of memoranda increases, observers' responses become increasingly imprecise and error-prone, yielding a high proportion of apparently random responses at higher set sizes. Zhang and Luck (2008) showed that this empirical pattern was well explained by a model that presumes a mixture of target-related responses and random guesses, suggesting that observers failed to store more than about three items in visual WM. Nevertheless, debate has continued on the basic question of whether observers ever fail to store relevant items in working memory, because the empirical pattern discovered by Zhang and Luck (2008) can also be modeled by presuming a relatively high proportion of "memories" that are so imprecise that they resemble random guesses (e.g., Bays et al., 2009; van den Berg et al., 2012, 2014). Thus, it has been difficult to find evidence that can distinguish between random guesses and responses based on extremely low fidelity memories.

Adam, Vogel and Awh (2017) offered new evidence that attempted to break this deadlock. They employed a whole report procedure in which observers recalled the color or orientation of all items presented, in whatever order they preferred. First, Adam et al. found that observers had a strong tendency to report items in descending order of memory quality. Thus, while the first response was highly precise, every response thereafter showed a monotonic rise in average error. Indeed, in trials with 6 memoranda, the fifth and sixth responses were best modeled with a completely random distribution of responses that had no relationship to the item to be recalled. This simple model (a uniform distribution with zero free parameters) provided a better fit than alternatives that attempted to explain performance without resorting to any guessing behaviors (such as with ultra-wide Von Mises distributions). Moreover, the prevalence of this guessing pattern precisely matched the prevalence with which observers *self-reported* that they were guessing as they performed these trials. Thus, these findings provided objective evidence for guessing that was strongly confirmed by subjects' own reports that they had no information to recall for approximately 3 out of 6 items.

Although we view the Adam et al. (2017) findings as strong evidence for guessing behaviors, they are dependent on two critical features of the data. First, the relatively pure guessing behaviors that we observed for the fifth and sixth responses depended on a non-obligatory tendency of observers to report the items in descending order of memory quality. Although we will present a close replication of this finding in the present work, the fact remains that this strong bias in response order is likely determined by the observer's response strategy rather than by any structural constraint in their recall of the memory items. In line with this, we are aware that at least one other study did not find the same "pure" evidence of guessing in the last responses of set size six trials (Oberauer, 2022), even though this study replicated the procedure used in Adam et al. (2017). One possibility is that the subjects in the Oberauer (2022) study simply chose a different strategy with regard to response order. This finding highlights the importance of methods that can enable a more confident interpretation of error distributions that contain a *mixture* of target-related and guess responses, rather than relying on the strong response bias observed by Adam et al. (2017).

Our approach to this challenge was inspired by a method that was introduced by Rouder, Thiele, Province, Cusumano, and Cowan (2014). In their experiments, subjects were required to recall the angle of a "gem" on the circumference of an open ring. The possible positions of the gem varied from -60° to +60° surrounding the vertical midline at the top of the ring. A key contribution of the Rouder et al. (2014) study was to document a distinctive pattern of responses that could be directly linked with guessing. They identified this guessing pattern by including "false probes" that asked subjects to recall the position of the gem on items that had never been presented. These false probes elicited a distinctive pattern of responses that were strongly biased towards 30° and +30°, the positions occupying the midpoint of the regions on the left and right sides of the ring. Given that no angle had been presented at all, this response pattern could be used as an explicit signature of guessing in this task. Critically, Rouder et al. (2014) also conceived of an analysis that could reveal this guessing pattern during trials in which subjects were asked to recall the gem position of real memory items. To this end, Rouder et al. (2014) plotted the responses with the reported angle of the gem on the y-axis, and the studied angle on the x-axis. For trials in which subjects knew the position of the gem, the resulting values were concentrated on the main diagonal of the scatterplot. By contrast, responses to false probes yielded horizontal "bands" that were centered at -30° and +30° on the y-axis. Critically, the same horizontal "guess bands" were observed when observers attempted to recall the gem position from trials with larger numbers of items, strongly implying that subjects were forced to guess on a subset of those trials. A key virtue of this approach is that the observed guess bands were clearly distinct from the pattern expected if responses were guided by extremely low precision memories; low precision memories should simply increase the width of the observed diagonal rather than yielding discrete horizontal bands that do not vary with the studied angle. Thus, Rouder et al. (2014) introduced an approach that enables a clear discrimination between random guessing states with zero information and extremely low precision memories that are nevertheless centered around the studied angle.

The present studies relied heavily on the insights from Rouder et al. (2014), but we also introduced some key changes to the procedure. First, we employed orientation stimuli in which a single "clock hand" could be presented anywhere in the 360-degree space around the circumference of a ring. We reasoned that this circular stimulus space would be more conducive to the kind of mixture modeling approaches introduced by Zhang and Luck (2008). In turn, this enabled a direct comparison between the prevalence of guessing inferred with the mixture modeling approach used by Zhang and Luck, and the prevalence of guessing as indicated by the presence of guess bands. Given that guess bands cannot be explained by reliance on extremely low precision memories, correspondence in the amount of guessing determined by the two approaches would provide strong leverage for interpreting the putative evidence for guessing in the Zhang and Luck analysis. Finally, in line with the Adam et al. (2017) whole-report approach, observers indicated with each response whether or not they thought they were guessing, enabling a test of whether objective evidence regarding the frequency of guessing would line up with the observers' meta-knowledge of whether an item had been stored.

To anticipate the results, the Rouder et al. (2014) analytic approach revealed guess bands in our procedure that were centered within each of the four quadrants defined by a colored background in each orientation stimulus. Interestingly, the location of the bias was largely driven by the background, such that rotating the angle of the background also yielded rotated guess bands. This finding suggests that the observed bias is apparently dependent on the specific stimulus context rather than determined by absolute angle or retinotopic position. Critically, the combined frequency of guess band responses and completely random responses was tightly correlated with the frequency with which observers reported that they were guessing, as well as with the guessing parameter obtained from the Zhang and Luck (2008) mixture modeling approach. These findings provide clear evidence for a large proportion of responses that are completely disconnected from the to-be-reported orientations, and easily distinguished from responses that are guided by imprecise memories. Thus, these results provide further evidence for guessing behaviors in visual working memory tasks.

Experiment 1
Methods
40 participants from the University of Oregon completed the experiment for class credit or payment ($8/hour). All participants reported normal or corrected-to-normal vision, and provided informed consent according to procedures by the University of Oregon institutional review board. The experiment code and data, and analysis code and output are openly accessible at https://osf.io/64rdq/.

Stimuli were generated using MATLAB (The MathWorks, Natick, MA) and PsychToolbox (Brainard, 1997; Pelli, 1997), and presented on a 17-inch CRT monitor with a 60 Hz refresh rate and a 1024 x 768 resolution. The viewing distance was approximately 60 cm. On each trial, participants were presented with six clock faces, each with a diameter of 2.7 degrees and depicting a different angle that was selected randomly, drawn in black on a mid-gray background for 200 msec. The spatial locations of each item were randomly generated within the region 8.5 degrees above and below fixation, and 11.9 degrees to the left and right of fixation. To minimize crowding, at least one item appeared in each quadrant of the screen and each item was separated by at least 0.65 degrees. In the ‘background’ condition, the clock faces appeared with a background of four red (RGB = 255, 102, 102) and green (RGB = 102, 255, 102) wedges centered on 45°, 135°, 225° and 315° from vertical, whereas in the ‘no background’ condition, the clock faces appeared without the background (see Figure 1). After a retention interval of 1 second with a blank screen, participants were shown a screen with circle outlines at the locations of the memoranda. Participants clicked within the outline of the item to select it, reporting whether they were confident or guessing by responding with either the left mouse button or the right mouse button respectively. Participants then clicked at an angle away from the center of the item to make their response. Participants could freely select the order of items they responded to, finishing the trial once they responded to all items. This was followed by a one-second inter-trial interval.

Participants completed four blocks of the orientation recall task without any colored background (‘standard’ condition), and then six blocks with the colored backgrounds (‘background’ condition). This condition order was not counterbalanced to avoid any carry-over effects that may occur if participants were to complete the ‘background’ condition first.

Each block had 20 trials each, making 80 trials in the no-background condition and 120 trials in the background condition per participant.

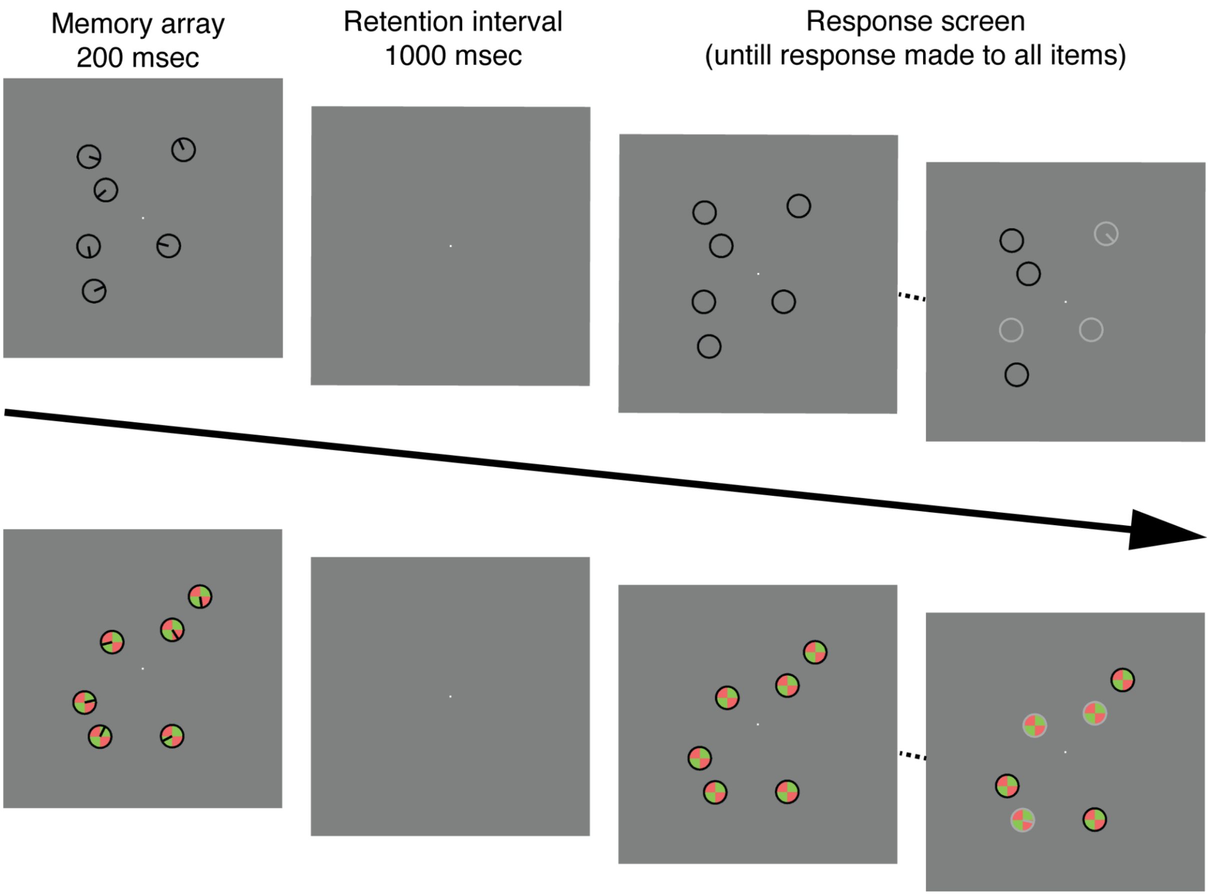

Figure 1. A schematic of an experiment trial from the ‘standard’ condition (top) and the ‘background’ condition (bottom). Six randomly selected angles are presented for 200 msec before a blank retention interval of 1000 msec. The locations of the memory array items were presented on the response screen. Participants clicked on which location they wanted to respond to and then moved the mouse cursor in the direction of the angle they wished to respond, and clicked down again to submit their response. Items with submitted responses were grayed out and the trial ended once all items had been responded to.

Analysis

We examined which of three models – the standard mixture model (Zhang & Luck, 2008), the variable precision and no-guessing model (van den Berg et al., 2014) and a parameter-free uniform distribution – best-fit each individual’s recall error distribution, separately for each response. We used the *MemFit* function from the MemToolbox package for MATLAB (Suchow et al., 2013) to conduct the formal model comparison, selecting the best-fitting model by determining which has the lowest the Bayesian Information Criterion (BIC) value.

We will also present the data as a scatterplot of the presented angle against the reported angle from the first to the sixth response. From visual inspection, one can see the main components of the dot distributions are a diagonal component (an accurate response based on memory) and horizontal bands (a response based on the colored quadrant backgrounds, and not based on any information held in memory). To verify this, we conducted a formal model comparison of the data for each of the six responses – both aggregated across all individuals, as well as each individual participants' data separately.

There were three possible components to our models - a memory response, a guess band response and a random response (see Figure 2). Firstly, memory responses will be centered on the presented angle but with some error dependent on the precision of the memory. We modeled this with a Von Mises distribution with its standard deviation as a free parameter (σ_{mem}). Secondly, participants may emit responses that are independent of the true angle, but biased towards the center of the colored quadrant backgrounds (a 'guess band' response). We modeled these responses with a combination of 4 Von Mises distributions with equal likelihood and standard deviation (σ_{bands}), each centered on the middle angle of a quadrant. Lastly, we examined whether there were also random responses that were not strongly biased by the background colors; these were modeled using a uniform distribution. We compared each possible combination of these components in our models, with $n - 1$ free parameters for the proportion of responses attributed to each component (P_{mem} and/or P_{bands}), where n is the number of components in a given model (Table 1).

Table 1

The components and number of free parameters of each of the compared models

Model	Components	Number of parameters
M1	Von Mises + Uniform (Zhang and Luck)	2: σ_{mem}, P_{mem}
M2	Guess Bands + Uniform	2: σ_{bands}, P_{bands}
M3	Von Mises + Guess Bands + Uniform	4: σ_{mem}, σ_{bands}, P_{mem}, P_{bands}
M4	Guess Bands	1: σ_{bands}
M5	Von Mises + Guess Bands	3: σ_{mem}, σ_{bands}, P_{mem}
M6	Von Mises	1: σ_{mem}

Note. The components of each model compared with the number of free parameters.

Using custom probability distribution functions, we conducted likelihood maximization to find the best-fitting parameters of each model to the aggregated response data, as well as to each participant's response data. Each instance of parameter estimation was repeated 100 times, starting from random parameter estimates, using the MATLAB *mle* function for a maximum of 10,000 iterations. To conduct comparison of models with varying numbers of parameters, we compute the Bayesian Information Criterion (BIC), which applies a penalty for having an additional number of free parameters, for each model. The best-fitting model was selected by finding the lowest BIC value.

We also compared the parameter estimates of guessing proportion of our model with the Zhang and Luck (2008) mixture model. Mixture modeling was conducted using the *MemToolbox* (Suchow et al., 2013). We also compared participant's proportion of self-reported guesses on each response with model estimates of guessing prevalence with linear mixed models using the *lme4* (Bates et al., 2007) package in R (R Core Development Team, 2013). The p-values for predictors were generated using the *lmerTest* package (Kuznetsova et al., 2017).

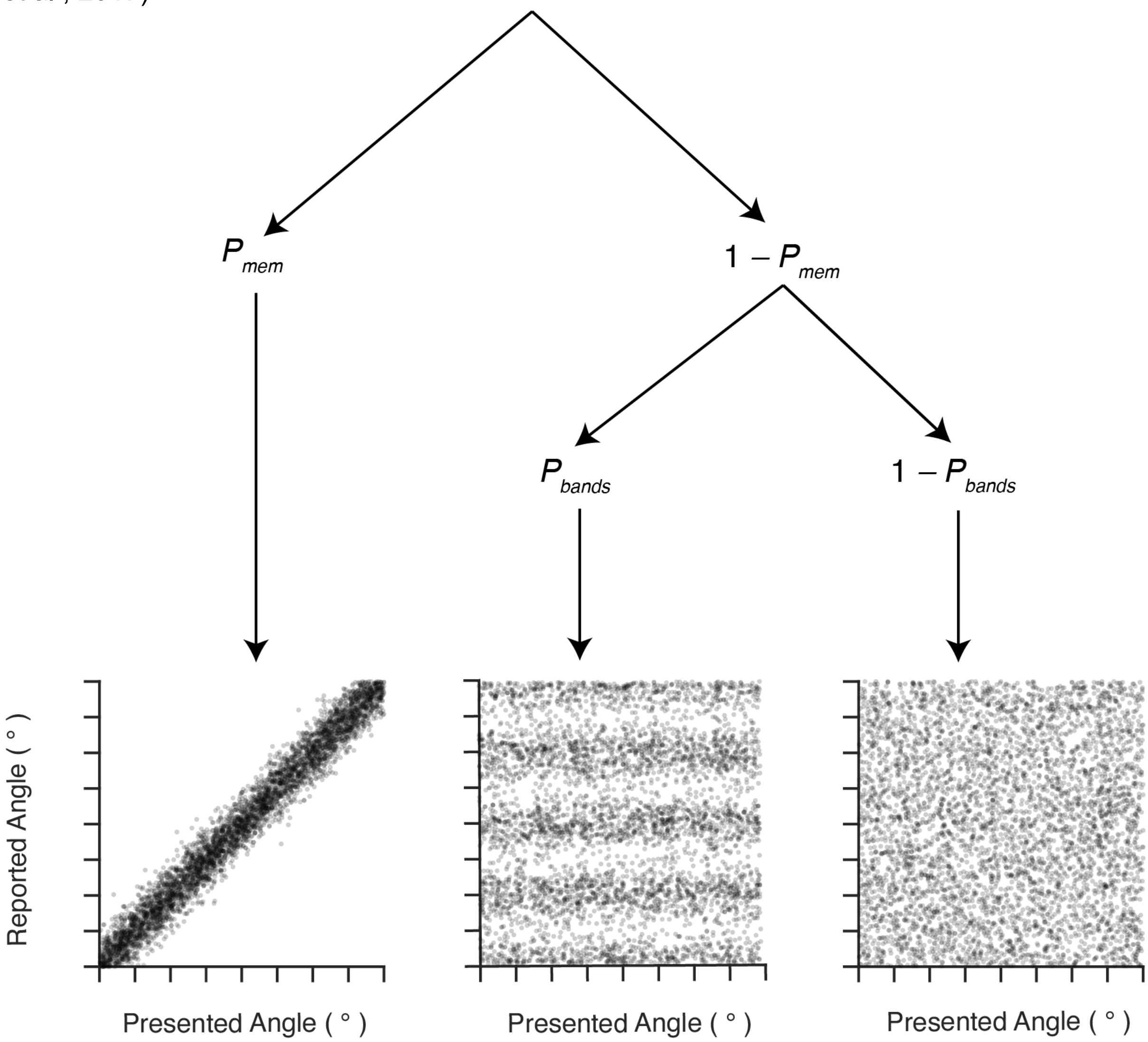

Figure 2. A schematic with simulated data patterns for all components in the complete model (M3). The prevalence of memory responses, 'guess band' responses and uniform responses are estimated with free parameters (P_{bands} and P_{mem}). The leftmost scatterplot is simulated data from a Von Mises distribution. The precision of memory representations, the width of the diagonal, is modeled with a free parameter (σ_{mem}). The middle scatterplot is simulated data from a 'guess band' component. The width of the horizontal bands is modeled with a free parameter (σ_{bands}). The rightmost scatterplot is simulated data from a uniform distribution, indicating a random response.

Results

The aggregated data from the background condition of Experiment 1 are plotted as error distributions (Figure 3a), and as scatterplots of the reported angle as a function of the presented angle (Figure 3b) and similarly from the no background condition (Figure 4). From visual inspection of the data in the background condition, the first three responses have a clear diagonal component indicating a substantial amount of accurate recall. However, in the last four responses, there is a clear banded pattern across the scatter plot, indicating participants were clicking towards the center of the quadrant backgrounds regardless of the presented angle. We refer to these responses as ‘guess bands’. The first three responses of the no background (standard) condition show the same diagonal component indicating accurate memory recall. In contrast to the background condition, there is not an obvious banded pattern across the scatter plots of the last four responses.

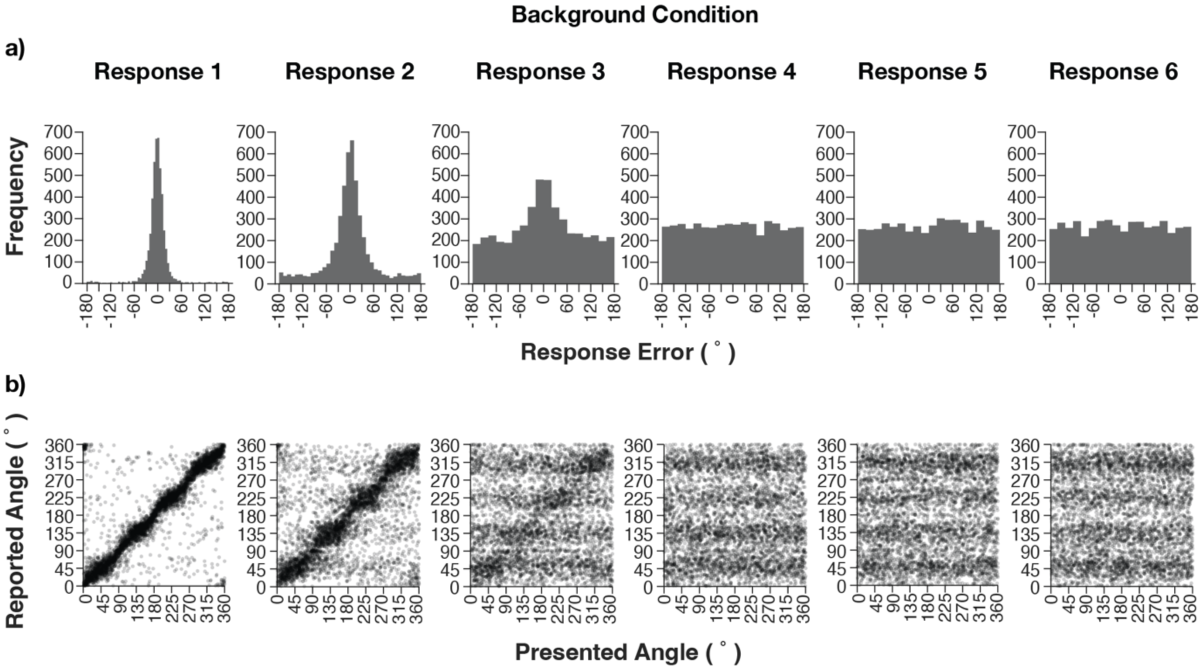

Figure 3. a) Response error distributions aggregated across all participants for each of the six responses in the background condition of Experiment 1. b) Scatterplots of the reported angle and the presented angle aggregated across all participants for each of the six responses in the background condition of Experiment 1. In early responses, there is a clear diagonal component produced by a substantial proportion of memory responses. In later responses, there is no longer a diagonal component but a banded pattern appearing at the center of the colored quadrant backgrounds (45˚, 135˚, 225˚ and 315˚).

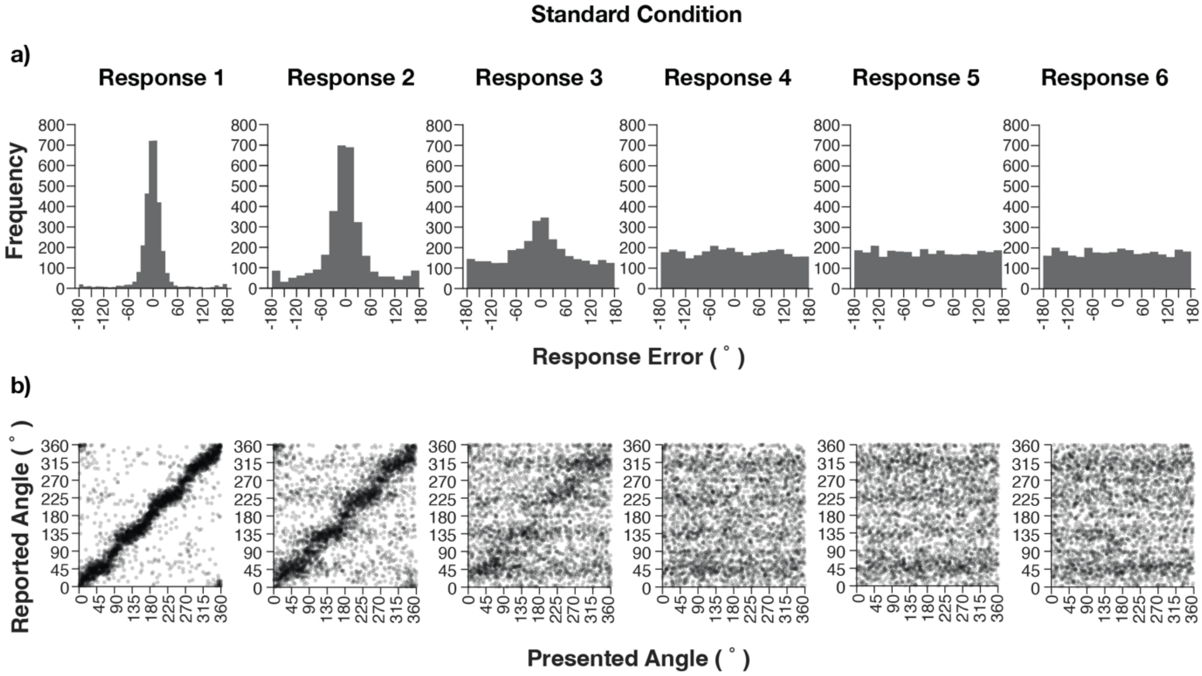

Figure 4. a) Response error distribution aggregated across all participants for each of the six responses in the 'no background' condition of Experiment 1. b) Scatterplots of the reported angle and the presented angle aggregated across all participants for each of the six responses. In early responses, there is a diagonal component produced by a substantial proportion of memory responses. In later responses, the diagonal component fades and responses appear to be mostly random.

<u>Formal Model Comparison of Individual Error Distributions</u>

We conducted a formal model comparison of each observer's error distributions separated by response, in both the standard condition (Table 2 and 3) and the background condition (Table 4 and 5). In the standard condition, the standard mixture model (Zhang & Luck, 2008) best-fit the majority of individual's first two responses of the background condition. For the third response, the uniform distribution was best-fitting model for most observers' errors, and the standard mixture model was the best-fitting for the remaining observers. Strikingly, for the last three responses, the best-fitting model for all observers (except for two observers on their fourth responses) was the parameter-free uniform distribution. The model comparisons had the same pattern of results in the background condition; the first two responses were best-fit by the standard mixture model, and the uniform distribution was best-fitting for all participants in their last three responses.

Table 2

Model comparison results for the individual error distributions in the standard condition across responses in Experiment 1

Response	Standard Mixture model	Variable Precision and no guessing model	Uniform distribution
1st	**39**	1	0
2nd	**39**	0	1
3rd	16	0	**24**
4th	2	0	**38**
5th	0	0	**40**
6th	0	0	**40**

Note. The number of participants for which each model was the best-fitting to their response data. The bolded number indicates the best-fitting model for the majority of participants for that response. While the early responses are best modeled by a standard mixture model for the majority of participants, the uniform distribution is the best-fitting from the 4th response onwards.

Table 3

BIC values of the formal model comparison for the individual error distributions in the standard condition across responses in Experiment 1

Response	Standard Mixture model	Variable Precision and no guessing model	Uniform distribution
1st	**741.50**	+5.22	+200.27
2nd	**867.19**	+6.98	+74.59
3rd	-2.91	+3.78	**941.78**
4th	+10.47	+16.83	**941.78**
5th	+11.48	+17.78	**941.78**
6th	+11.12	+17.57	**941.78**

Note. The bolded number indicates the BIC value for the best-fitting model for the majority of participants for that response. The remaining values in each row are the mean difference in BIC values to the best-fitting model.

Table 4
Model comparison results for the individual error distributions in the background condition across responses in Experiment 1

Response	Standard Mixture model	Variable Precision and no guessing model	Uniform distribution
1st	**35**	5	0
2nd	**36**	1	3
3rd	**20**	0	20
4th	0	0	**40**
5th	0	0	**40**
6th	0	0	**40**

Note. The number of participants for which each model was the best-fitting to their response data. The bolded number indicates the best-fitting model for the majority of participants for that response. While the early responses are best modeled by a standard mixture model for the majority of participants, the uniform distribution is the best-fitting from the 4th response onwards.

Table 5
BIC values of the formal model comparison for the individual error distributions in the background condition across responses in Experiment 1

Response	Standard Mixture model	Variable Precision and no guessing model	Uniform distribution
1st	**1067.7**	+4.06	+344.98
2nd	**1276.3**	+5.61	+136.37
3rd	**1406.2**	+6.80	+6.47
4th	+11.94	+18.59	**1412.7**
5th	+12.21	+19.10	**1412.7**
6th	+11.95	+19.09	**1412.7**

Note. The bolded number indicates the BIC value for the best-fitting model for the majority of participants for that response. The remaining values in each row are the mean difference in BIC values to the best-fitting model.

Formal Model Comparison of Aggregated Stimulus Against Response Data
For the first three responses of the background condition, the best-fitting model is M5 (Von Mises + Guess Bands) closely followed by M3 (Von Mises + Guess Bands + Uniform) (see Table 6 for BIC values of all models for each response). The best parameter estimates of the

complete model from maximum likelihood estimation, and the calculated proportion of responses from each component (memory, guess bands and uniform) can be seen in Table 7. The proportion of responses attributed to a memory representation decreased from first to third response before being essentially non-existent in the last three responses. The proportion of responses attributed to the guess band component from the quadrants is substantial after the second response, converging with visual inspection of the scatter plots (Figure 3).

Table 6

Results of model comparison on aggregated data across responses in Experiment 1

BIC	**M1**	**M2**	**M3**	**M4**	**M5**	**M6**
1st	+66	+13518	+9	+13510	**42623**	+2780
2nd	+217	+5085	+9	+5077	**50994**	+1332
3rd	+449	+436	+3	+432	**55563**	+4761
4th	+412	+105	**55946**	+109	+4	+7656
5th	+450	+191	**55803**	+230	+38	+7723
6th	+376	+272	**55769**	+284	+12	+7775

Note. The Bayesian Information Criterion (BIC) values for the best-fitting iteration of each model to the aggregated data of each of the six responses. A lower BIC value indicates a better fit. The lowest BIC value for each whole-report response is bolded. For the first three responses, the best-fitting model is the Von Mises + Guess Bands (M5) and for the last three responses, the best-fitting model is the Von Mises + Guess Bands + Uniform (M3).

Table 7

Parameter estimates from the best-fitting complete model for each response in Experiment 1

Response	σ_{mem}	σ_{bands}	P_{mem}	P_{bands}	P_{unif}
1st	14.3225	21.6940	90.59% ± 0.57%	9.41% ± 1.15%	0% ± 0.58%
2nd	23.7556	22.4566	66.03% ± 1.68%	33.97% ± 2.20%	0% ± 0.52%
3rd	30.8256	17.7327	20.37% ± 0.63%	46.64% ± 12.16%	32.99% ± 11.53%
4th	0	16.3337	0.19% ± 0.09%	41.96% ± 8.29%	57.85% ± 8.20%
5th	0	13.4376	0.30% ± 0.12%	35.78% ± 4.53%	63.92% ± 4.41%
6th	0	15.2980	0.39% ± 0.12%	39.12% ± 6.25%	60.49% ± 6.13%

Note. The estimates and 95% confidence interval for the prevalence of memory, guess bands and uniform components shown here were calculated from the best-fitting model parameters. The first three responses contain a substantial proportion of memory responses, whereas the last three responses contain a mixture of random guesses and strategic responses based on the quadrant backgrounds.

One possible concern is whether the monotonic decline we observed across the six responses could be explained by *output interference*. Fortunately, Adam et al. (2017) examined this issue using almost identical orientation stimuli (minus the colored backgrounds used in the current work) and the same whole report procedure. To estimate the effect of output interference, Adam et al. (2017) measured performance in a computer-guided whole report task in which observers were no longer free to choose the order of responses. Although the computer-guided procedure revealed a detectable decline across the six responses, the slope of the decline was only 16% of the slope that was observed when observers were free to choose response order themselves. Thus, output interference cannot explain 84% of the decline, and cannot be the reason why less than 1% of clicks had information about the target after the third response.

Formal Model Comparison of Individual Stimulus Against Response Data

We also conducted a model comparison of each individual's data in the background condition, separated by each response. It should be noted that this significantly reduced the amount of data (120 trials per click response) for maximum likelihood estimation. In the early responses, the best-fitting model to most individuals' response data was either the Von Mises + Uniform (M1) or Von Mises + Guess Bands (M5), before a notable switch to the Guess Bands only (M4) from the 4th response onwards (Table 8). This indicates that while there is some variation between individuals, the substantial majority do anchor to the background quadrants with their guesses in later responses.

Table 8

Model comparison for the individual data across responses in Experiment 1

Response	M1	M2	M3	M4	M5	M6
1st	**28**	-	-	-	10	2
2nd	**19**	-	1	2	18	-
3rd	**14**	-	1	2	13	-
4th	6	-	-	**30**	4	-
5th	5	2	2	**25**	6	-
6th	6	1	2	**23**	8	-

Note. The number of participants for which each model was the best-fitting to their response data. The bolded number indicates the best-fitting model for the majority of participants for that response. While the early responses are best modeled by either a Von Mises + Uniform (M1) or Von Mises + Guess Bands (M5) for the majority of participants, there is a notable switch to the Guess Bands only (M4) from the 4th response onwards.

Comparison to Standard Mixture Modeling Approaches

We compared the results of our analyses to those from the mixture modeling approach employed by Zhang and Luck (2008) to argue for the presence of random guessing responses in continuous report tasks. Typically, only a single item of the memory array is probed on each trial, and mixture modeling is performed on the error distribution of all trials.

Here, we fit our general model (Von Mises + Guess Bands + Uniform) as well as the Zhang and Luck (2008) mixture model to the individual data aggregated across all responses on each trial (the same population that the single probe is sampling from).

When the prevalence of guess bands and uniform responses (from the Rouder-style analysis) were considered separately, neither one correlated with the prevalence of guessing as estimated using the Zhang and Luck (2008) mixture model (guess bands: $r = 0.06$, $p = 0.70$; uniform: $r = 0.22$, $p = 0.17$). However, when the combined frequency of guess bands and uniform responses was examined, there was a powerful correlation with the prevalence of guessing according to the Zhang and Luck analysis ($r = 0.95$, $p < .001$). Further, we conducted a linear mixed model predicting estimates of guess prevalence (i.e., combined prevalence of guess bands and uniform) with observers' proportion of self-reported guesses across responses, allowing a random intercept for each participant (Figure 5). The proportion of self-reported guesses was a strong predictor of guessing prevalence ($\beta = 0.903$, $SE_b = 0.032$, $t(213.64) = 28.33$, $p < .001$). Thus, although subjects vary in the proportion of guesses that are strongly biased by the background, guess band responses can be directly linked with the uniform guessing distribution observed with the Zhang and Luck (2008) mixture model, as well as with observers' own reports of how often they were guessing.

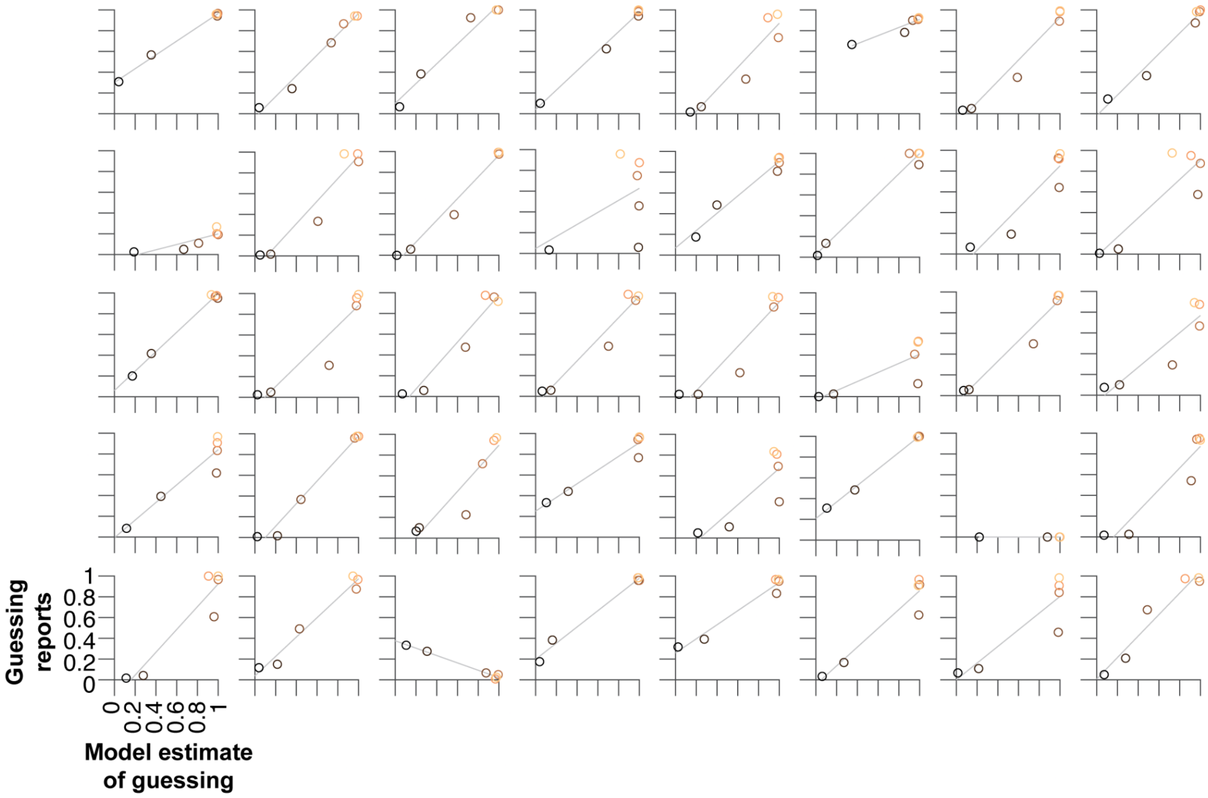

Figure 5. Scatterplots of the relationship between model estimates of the prevalence of guessing and participants' self-reports of guessing on each response click in Experiment 1.

Experiment 2

Experiment 2 had two primary purposes. First, we sought to replicate the patterns observed in Experiment 1. Second, we examined whether the bias towards the center of the colored quadrants was based on the absolute angle within the orientation stimulus (e.g., as might be expected if the bias were related to the oblique effect), or whether the bias was instead dependent on the specific stimulus that was presented. To this end, we rotated the quadrants by 45°, shifting the centers of the quadrants. To anticipate the results, the guess bands followed the position of the colored backgrounds, suggesting that they are determined by the specific backgrounds within each orientation stimulus, not a bias towards a specific set of orientation values.

Methods
The methods were identical to Experiment 1 except for the following details. 30 participants from the University of Oregon completed the experiment. Participants completed 8 blocks of 20 trials with the colored quadrant backgrounds. The colored backgrounds were rotated such that they were centered at 0°, 90°, 180° and 270° from vertical. There were no trials without the colored background.

Results
We replicated the results from Experiment 1. The aggregated data of Experiment 2 are plotted as error distributions (Figure 6a), and as scatterplots of the reported angle as a function of the presented angle (Figure 6b). From visual inspection, the first three responses have a clear diagonal component indicating a substantial amount of accurate recall. As in Experiment 1, in the last four responses, there is an obvious band pattern across the scatter plot, indicating participants were clicking towards the center of the quadrant backgrounds (0°, 90°, 180°, 270° from vertical) regardless of the presented angle.

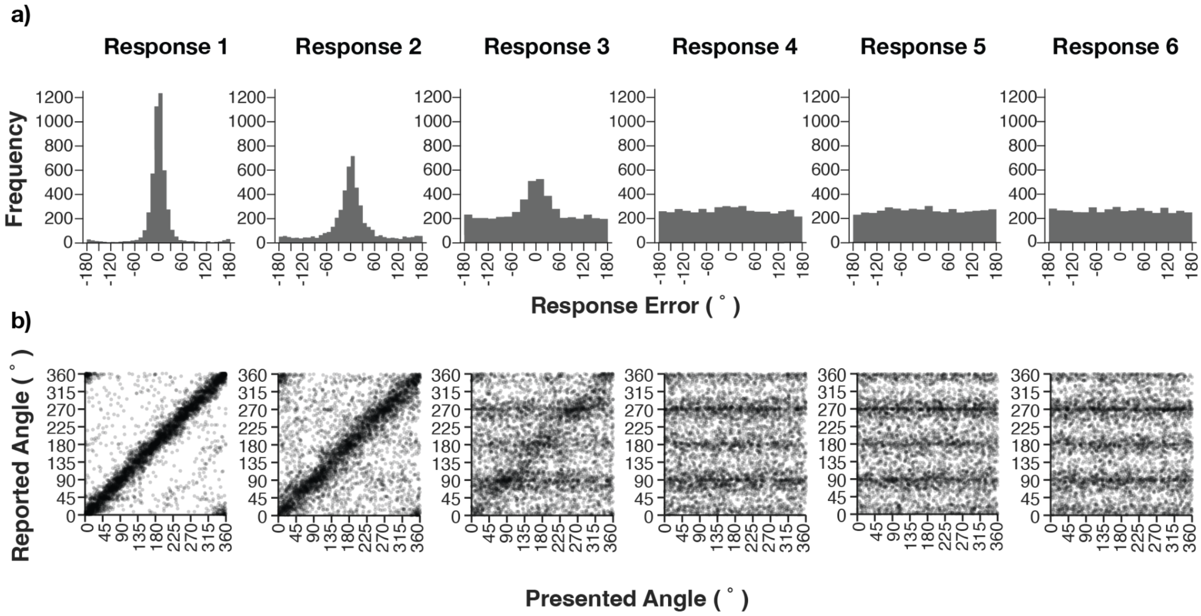

Figure 6. a) Response error distributions aggregated across all participants for each of the six responses in the background condition of Experiment 2. b) Scatterplots of the reported angle and the presented angle aggregated across all participants for each of the six responses in the background condition of Experiment 2.

<u>Formal Model Comparison of the Individual Error Distributions</u>
As in Experiment 1, we analyzed individual's error distributions for each response with a formal model comparison between the standard mixture model, variable precision and no-guessing model, and a uniform distribution (Tables 9 and 10). For the first two responses, the best-fitting model for the majority of observers was the standard mixture model. It was a

mixture for the third response, where the best-fitting model was the uniform distribution for the majority of observers' error distributions and the standard mixture model for the remaining observers. But again, strikingly, for the last three responses, all observers' error distributions (save for the fourth response of three observers) was best-fit by the uniform distribution.

Table 9

Model comparison results for the individual error distributions in Experiment 2

Response	Standard Mixture model	Variable Precision and no guessing model	Uniform distribution
1st	**37**	3	0
2nd	**39**	1	0
3rd	15	0	**25**
4th	3	0	**37**
5th	0	0	**40**
6th	0	0	**40**

Note. The number of participants for which each model was the best-fitting to their response data. The bolded number indicates the best-fitting model for the majority of participants for that response. While the early responses are best modeled by a standard mixture model for the majority of participants, the uniform distribution is the best-fitting from the 4th response onwards.

Table 10

BIC values of the formal model comparison for the individual error distributions across responses in Experiment 2

Response	Standard Mixture model	Variable Precision and no guessing model	Uniform distribution
1st	**1444.2**	+4.08	+430.67
2nd	**1695.0**	+5.77	+179.89
3rd	-25.72	-19.01	**1874.9**
4th	+10.14	+17.31	**1874.9**
5th	+11.81	+19.08	**1874.9**
6th	+12.66	+19.94	**1874.9**

Note. The bolded number indicates the BIC value for the best-fitting model for the majority of participants for that response. The remaining values in each row are the mean difference in BIC values to the best-fitting model.

Formal Model Comparison of the Aggregated Stimulus Against Response Data

We conducted a formal model comparison of the aggregated data in Experiment 2. This was the same comparison as in Experiment 1 (see Table 6), noting that the Von Mises distributions modeling the guess bands were centered at 0°, 90°, 180, 270° to reflect the rotation of the colored quadrant backgrounds. The best-fitting model to the first response was M1 (Von Mises + Uniform), to the second response was M5 (Von Mises + Guess Bands) and then M3 (Von Mises + Guess Bands + Uniform) for the remaining responses (Table 11), closely mirroring the results of Experiment 1 (Table 3).

The best-fitting parameter estimates of the complete model (Von Mises + Guess Bands + Uniform) are reported in Table 12. According to this model, the estimated prevalence of memory responses gradually declined from the 1st to the 3rd response before reaching essentially zero from the fourth response onward. The prevalence of attributed to the 'guess bands' component is substantial from the third response onward, similarly to Experiment 1.

Table 11

Results of model comparison on aggregated data across responses in Experiment 2

BIC	M1	M2	M3	M4	M5	M6
1st	**43650**	+12476	+6	+12535	+8	+3028
2nd	+22	+5069	+1	+5069	**51075**	+1409
3rd	+521	+556	**55107**	+572	+15	+4712
4th	+1128	+238	**55120**	+282	+41	+7720
5th	+1039	+123	**55039**	+190	+68	+8053
6th	+1072	+146	**54968**	+219	+71	+8312

Note. The Bayesian Information Criterion (BIC) values for the best-fitting iteration of each model to the aggregated data of each of the six responses in Experiment 2. A lower BIC value indicates a better fit. The BIC value of the best-fitting model of each response has been bolded. For the first response, the best-fitting model is the Standard Mixture Model (M1). For the second response, the best-fitting model is Von Mises + Guess Bands (M5). For the third to sixth response, the best-fitting model is the Von Mises + Guess Bands + Uniform (M3).

Table 12
Parameter estimates from the best-fitting complete model for each response in Experiment 2

Response	σ_{mem}	σ_{bands}	P_{mem}	P_{bands}	P_{unif}
1st	14.3225	21.6940	87.84% ± 0.00%	0.64% ± 0.00%	11.52% ± 0.00%
2nd	23.7556	22.4566	64.13% ± 1.18%	2.08% ± 0.90%	33.79% ± 2.08%
3rd	30.8256	17.7327	21.07% ± 0.61%	37.26% ± 6.25%	41.67% ± 5.65%
4th	0	16.3337	0.31% ± 0.11%	48.10% ± 6.02%	51.59% ± 5.91%
5th	0	13.4376	0.21% ± 0.11%	48.70% ± 4.70%	51.09% ± 4.58%
6th	0	15.2980	0.25% ± 0.11%	47.22% ± 4.35%	52.53% ± 4.24%

Note. Estimates with the range of the approximate 95% confidence intervals for the proportion of memory, guess bands and random responses were calculated from the parameter estimates of the best-fitting complete model to the aggregated data of Experiment 2. The first three responses contain a substantial proportion of memory responses, whereas the last three responses contain a mixture of random guesses and strategic responses based on the quadrant backgrounds.

Formal Model Comparison of Individual Stimulus Against Response Data

In addition, we conducted a model comparison of each individual's data, separated by each response in the whole-report (1st to 6th). It should be noted that this significantly reduced the amount of data (160 trials) for maximum likelihood estimation. For the first two responses, the best-fitting model to most participants' data was the M1 (Von Mises + Uniform) model. For the third response, it was the M5 (Von Mises + Guess Bands) model, and for the last three responses, the M4 (Guess Bands only) model was the best-fitting for the majority of participants (see Table 13).

Table 13
Model comparison for the individual data across responses in Experiment 2

Response	M1	M2	M3	M4	M5	M6
1st	**23**	-	1	-	4	2
2nd	**17**	-	3	-	10	-
3rd	4	4	5	7	**10**	-
4th	4	7	5	**9**	5	-
5th	5	**11**	1	**11**	2	-
6th	1	5	3	**16**	5	-

Note. The number of participants for which each model was the best-fitting to their response data. While the early responses are best modeled by either a Von Mises + Uniform (M1) or Von Mises + Guess Bands (M5) for the majority of participants, there is a notable switch to the Guess Bands only (M4) model from the 4th response onwards.

Comparison to Standard Mixture Model Approaches

As in Experiment 1, we fit our general model (Von Mises + Guess Bands + Uniform) as well as the Zhang and Luck (2008) mixture model to the individual data aggregated across the click responses on each trial (the population that the single probe is sampling from). Considered in isolation, neither the prevalence of responses that were random (associated with a uniform distribution) nor the prevalence of guess bands correlated with the guess proportion estimates from the Zhang and Luck (2008) mixture model (uniform: $r = 0.30$, $p = 0.11$; guess band: $r = -0.06$, $p = 0.75$). However, the combined frequency of guess bands and random uniform responses was tightly correlated with the guess proportion estimates from the Zhang and Luck mixture model ($r = 0.96$, $p < .0001$; Figure 7). Further, we conducted a linear mixed model predicting prevalence of guesses with observers' proportion of self-reported guesses on each response, allowing a random intercept for each participant. Proportion of self-reported guesses was a significant predictor of the prevalence of guessing (guess bands plus uniform) ($\beta = 0.847$, $SE_b = 0.038$, $t(165.47) = 22.35$, $p < .001$). In line with the results of Experiment 1, these finding show that the 'guess band' responses are coming from the same population as the uniform distribution in the Zhang and Luck model, although subjects vary in the relative proportions of guess band and uniform responses. These findings provide clear evidence that a substantial proportion of the responses identified as guesses with the Zhang and Luck (2008) analysis can be attributed to a "guess band" pattern that is completely disconnected from the probed item's orientation.

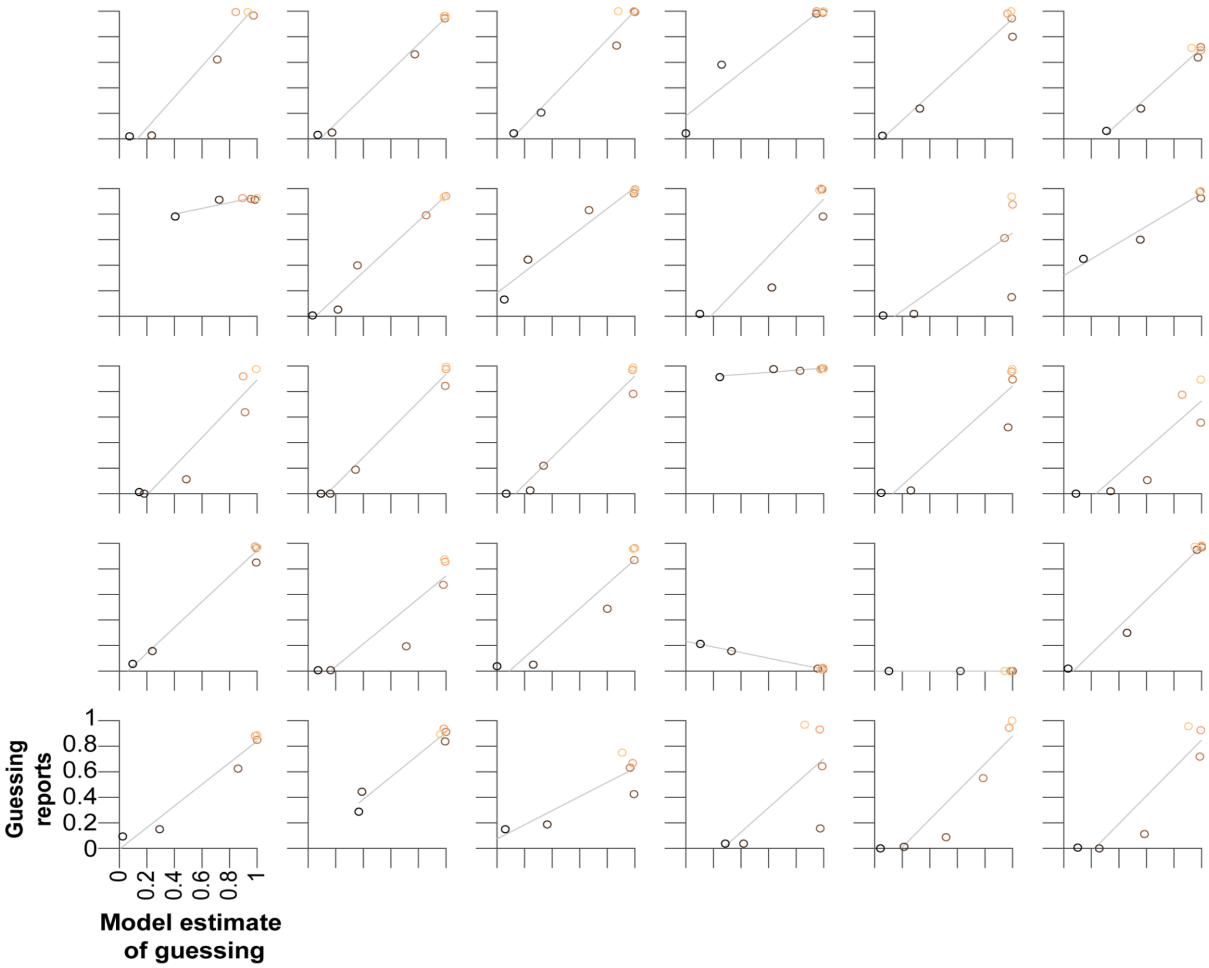

Figure 7. Scatterplots of the relationship between model estimates of the prevalence of guessing and participants' self-reports of guessing on each response click in Experiment 2.

Discussion

The central question in the current work is whether or not observers ever fail to store items that they wish to encode into working memory. If such storage failures occur, then there is a clear prediction that observers will be forced to guess when cued to recall that item. Although past work has presented evidence for such guessing responses (e.g. Adam et al., 2017; Nosofsky & Donkin, 2016; Nosofsky & Gold, 2018; Rouder et al., 2008; Zhang & Luck, 2008), the same empirical patterns can also be fit by models that eschew guessing but allow for "memories" so imprecise that they cannot be distinguished from storage failures (e.g. Bays & Husain, 2008; Fougnie et al., 2012; Schurgin et al., 2020; van den Berg et al., 2012). This theoretical impasse motivates analytic approaches that can distinguish between storage failures and extremely low fidelity memories. In this context, Rouder et al. (2014) demonstrated that observers have a strong tendency to cluster guess responses away from salient categorical boundaries in orientation space, without any link between these responses and any studied angle. The key virtue of this approach is that these "guess bands" can be easily distinguished from responses that are guided by very low precision memories, offering a potential resolution to the impasse described above.

In the present work, we offer a conceptual replication of the Rouder et al. (2014) findings, while also providing two key extensions. First, we employed orientation stimuli with a circular stimulus space that afforded a direct comparison with the mixture model analyses that

Zhang and Luck (2008) used to provide initial evidence for random guessing in a visual working memory task. Indeed, estimates of guessing frequency from the Zhang and Luck mixture modeling approach were tightly correlated with the combined frequency of guess bands and random responses in our primary analysis. Thus, our findings draw a direct link between the evidence for guessing from the Zhang and Luck (2008) approach and the guess bands observed using the Rouder et al. (2014) analysis. Given that the guess bands cannot be explained as responses guided by low precision memories, the present findings show that a substantial proportion of the uniform guessing distributions observed with the Zhang and Luck (2008) approach result from true guesses rather than low precision memories. The second extension provided by the current study was to validate our analytic estimates of guessing with observers' self-reports of guessing. In line with the prior work from Adam et al. (2017), there was a tight correlation – at the single subject level – between the combined frequency of guess bands and random responses in our analysis, and the frequency with which observers reported that they were guessing. Thus, these guess bands corroborate Zhang and Luck's (2008) interpretation of the uniform distributions observed during mixture modeling, and fall in line with observer's self-reports that they are guessing for approximately 3 items out of 6.

Our formal model comparisons match those in a similar extension of Rouder et al. (2014) conducted by Nosofsky and Gold (2018). On a delayed color recall task, Nosofsky and Gold (2018) biased participants to respond with one of sixteen possible choices on a circular color wheel by giving it potential of a high reward. In their task paradigm, a mixed-state model (one that contains an item limit and guessing beyond this limit) predicts accurate recall when the participant had memory of the presented color, but if the participant had no memory, they would be compelled to guess with the high-reward choice. On the other hand, a variable-precision and no guessing model predicts a range of responses between the presented color and high-reward choice, resulting from memories of various intermediary precision. Much like our results, their formal model comparisons showed their observed data was better-explained by a mixed-state model than by a variable-precision model that denied guessing.

One possibility that we did not explore is the existence of coarse categorical working memory representations (Bae et al., 2015; Hardman et al., 2017; Rouder et al., 2014; Souza et al., 2021). An intriguing possibility is that memory representations could be continuous or categorical in mixed proportions (Hardman et al., 2017). In the present study, the individual may remember the quadrant of the presented orientation but not the orientation itself for a proportion of responses. We did not include such a component in our formal model comparisons because of the overlap between the 'guess bands' component and what would be the step-like component for these categorical memories, and we lack the trials to be able to reliably distinguish these. From visual inspection of the scatter plots of Experiment 1 (Figure 3), we do somewhat observe a step-like pattern in the first three responses, but for the remaining responses, we see a clear horizontal banded pattern, indicating guesses appear across the range of presented angles rather than being contained within a remembered quadrant category in a step-like manner. The step-like component is less clear in the scatter plots of Experiment 2 (Figure 6) perhaps because the center of the colored quadrants match the cardinal orientations (Pratte et al., 2017).

We note that a recent article by Oberauer (2022) did not replicate the empirical pattern reported here and by Adam et al. (2017), whereby guesses (uniform distributions) are observed for the late response items. Oberauer (2022) reported that about half of their participants showed evidence consistent with zero-information states for the later responses, whereas the other half did not. Given that we have observed this empirical pattern in numerous independent samples of subjects (i.e., two in the Adam et al. (2017), two in the present study, and several additional experiments in an in-progress manuscript), we are unsure of the cause for this discrepancy. Oberauer raises the possibility that individuals may choose to guess despite having representations of some nontrivial precision for all array items, but he also noted that observers in his study may not have been as inclined to report the best remembered items first. Although extant data cannot discriminate between these possibilities, it is striking that the central question has shifted towards *why* observers are guessing, rather than *whether* they are guessing. Nevertheless, the discrepancy between Oberauer's findings and our own highlights the importance of methods that enable strong inferences of guessing without relying on the purely uniform distributions that we have documented in the present work. Even within distributions that contain mixtures of target-related and guessing responses, guess bands reveal a large proportion of responses that are wholly disconnected from the tested items without being confusable with response patterns guided by extremely low precision memories.

Because evidence for guessing behaviors is the central issue in this paper, it is appropriate to discuss our working definition of "guessing". Our use of the term refers to any response that is wholly unguided by information about the to-be-reported item. An alternate view of guessing is one in which "guessing" entails the use of a completely independent process for generating responses than the process that is used to report true memories. From this perspective, even if an observer has no information about the item to be recalled, the response is not considered to be a "guess" if the same retrieval operations are deployed as when a true memory of the item is available (Bays & Taylor, 2018; Schneegans et al., 2020; Schurgin et al., 2020). We disagree with this definition, because it obscures the fact that observers had no available memory of the reported attribute during recall. That said, this semantic debate about how guesses should be defined does not undermine the strong evidence for storage failures in visual working memory (the "guess bands"), an empirical pattern that disconfirms any model that posits non-zero memory guidance for all responses, regardless of the number of items to be stored (e.g. Bays & Husain, 2008; Fougnie et al., 2012; Schurgin et al., 2020; van den Berg et al., 2012).

That said, it is possible to conceive of strong modifications to continuous resource models that would enable tight fits with the response patterns we observed in the present work, at least when all responses are aggregated into a single distribution. For example, consider a continuous resource model in which there is always non-zero information about the target, but as the quality of the memory declines, the influence of an underlying prior increases. In this case, the observed 'guess bands' in the present study could be explained via an existing prior that contains bias towards the center of the quadrants, and an extremely fuzzy memory representation that cannot be detected. In this case, the modified continuous resource model could predict the emergence of guess bands that are wholly disconnected from the probed feature value. Here, it is critical to note that this modified continuous resource model predicts a large proportion of responses that are completely unrelated to the "remembered" item, an empirical pattern that fits our working definition of guessing. Thus, while this strong revision of continuous resource models would enable a tight fit with our observations, it requires an embrace of a "mixed state" model in which target-related responses are accompanied by a large proportion of responses with no discernable relationship to the probed value.

We maintain a preference for a guessing interpretation of our results for several reasons. Firstly, in both experiments, the recall error distributions for the last three responses (fourth to sixth) were better modelled by a parameter-free uniform distribution rather than the standard mixture model (Zhang & Luck, 2008) or the variable precision and no guessing model (van den Berg et al., 2014). Thus, because of the strong bias for observers to report the best remembered items first, a pure guessing model was superior to models that presume extremely low precision memories. Secondly, Adam et al. (2017) demonstrated that such a variable precision model may in fact be mimicking guessing responses. Simulations by Adam et al. (2017) indicated that it would require at least a million noise-free samples to discriminate between a uniform distribution and a preposterously low quality memory (modeled with a von Mises distribution with a standard deviation of 193 degrees). Reservations about the falsifiability of such a model aside (Nosofsky & Donkin, 2016; van den Berg et al., 2014), Adam et al. (2017) found in their experimental data (using a near-identical set-up to those in our studies, whole-report with clock face stimuli) that the estimated prevalence of these preposterously low quality memories was tightly correlated with guess rate estimates from the standard mixture models (Zhang & Luck, 2008). Lastly, as mentioned above, observers' self-reports indicate they had no knowledge of the presented stimulus at a proportion that matches our model estimates of the prevalence of guessing, adding credence to that interpretation. Thus, while strong modifications of a continuous resource model could yield strong fits of these findings, models acknowledging the possibility of storage failures provide a simpler explanation that resonates with the observers' own reports of when they have no information about the probed values.

In conclusion, we found clear evidence for guessing behavior that could not have been guided by low precision memories. These responses were prevalent in the later responses of our whole-report task, and could be directly linked with observers' self-reports of guessing. Thus, these findings provide further evidence for the guessing behaviors that are predicted by item-limit accounts of working memory capacity.

Open Practices Statement

The data and code for all experiments are available on the Open Science Framework at https://osf.io/64rdq/.

Declarations

This work was funded by grants from the National Institutes of Health R01 MH087214 and National Science Foundation 2104630. The authors have no competing interests to declare.

Acknowledgements

The contributions of the authors according to CRediT are as follows:
Conceptualization: J.J.F., K.C.S.A., and E.A. **Data curation:** W.X.Q.N. **Formal analysis:** W.X.Q.N., J.J.F., K.C.S.A., and E.A. **Funding acquisition:** E.A. **Investigation:** J.J.F. and K.C.S.A. **Methodology:** W.X.Q.N., J.J.F., and K.C.S.A. **Project administration:** W.X.Q.N. **Software:** W.X.Q.N., J.J.F., and K.C.S.A. **Supervision:** E.A. **Validation:** W.X.Q.N. **Visualization:** W.X.Q.N. and J.J.F. **Writing - original draft:** W.X.Q.N., J.J.F., K.C.S.A., and E.A. **Writing - review & editing:** W.X.Q.N., J.J.F., K.C.S.A., and E.A.

References

Adam, K. C. S., Vogel, E. K., & Awh, E. (2017). Clear evidence for item limits in visual working memory. *Cognitive Psychology*, *97*, 79–97. https://doi.org/10.1016/j.cogpsych.2017.07.001
Bae, G.-Y., Olkkonen, M., Allred, S. R., & Flombaum, J. I. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General*, *144*(4), 744–763. https://doi.org/10.1037/xge0000076
Bates, D., Sarkar, D., Bates, M. D., & Matrix, L. (2007). The lme4 package. *R Package Version*, *2*(1), 74.
Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision*, *9*(10), 7–7. https://doi.org/10.1167/9.10.7
Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, *321*(5890), 851–854.
Bays, P. M., & Taylor, R. (2018). A neural model of retrospective attention in visual working memory. *Cognitive Psychology*, *100*, 43–52. https://doi.org/10.1016/j.cogpsych.2017.12.001
Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.
Fougnie, D., Suchow, J. W., & Alvarez, G. A. (2012). Variability in the quality of visual working memory. *Nature Communications*, *3*(1), 1229. https://doi.org/10.1038/ncomms2237
Hardman, K. O., Vergauwe, E., & Ricker, T. J. (2017). Categorical Working Memory Representations are used in Delayed Estimation of Continuous Colors. *Journal of Experimental Psychology. Human Perception and Performance*, *43*(1), 30–54. https://doi.org/10.1037/xhp0000290
Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*, 1–26. https://doi.org/10.18637/jss.v082.i13
Nosofsky, R. M., & Donkin, C. (2016). Response-time evidence for mixed memory states in a sequential-presentation change-detection task. *Cognitive Psychology*, *84*, 31–62. https://doi.org/10.1016/j.cogpsych.2015.11.001
Nosofsky, R. M., & Gold, J. M. (2018). Biased guessing in a complete-identification visual-working-memory task: Further evidence for mixed-state models. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(4), 603.
Oberauer, K. (2022). Little Support for Discrete Item Limits in Visual Working Memory. *Psychological Science*, *33*(7), 1128–1142. https://doi.org/10.1177/09567976211068045
Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442.
Pratte, M. S., Park, Y. E., Rademaker, R. L., & Tong, F. (2017). Accounting for stimulus-specific variation in precision reveals a discrete capacity limit in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(1), 6–17. https://doi.org/10.1037/xhp0000302
R Core Development Team. (2013). *R: A language and environment for statistical computing*.
Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(16), 5975–5979. https://doi.org/10.1073/pnas.0711295105
Rouder, J. N., Thiele, J. E., Province, J. M., Cusumano, M., & Cowan, N. (2014). *The Evidence for A Guessing State in Working-Memory Judgments*.

https://web.archive.org/web/20161003063744/http://pcl.missouri.edu/sites/default/files/p_0.pdf
Schneegans, S., Taylor, R., & Bays, P. M. (2020). Stochastic sampling provides a unifying account of visual working memory limits. *Proceedings of the National Academy of Sciences*, *117*(34), 20959–20968. https://doi.org/10.1073/pnas.2004306117
Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2020). Psychophysical scaling reveals a unified theory of visual memory strength. *Nature Human Behaviour*, *4*(11), 1156–1172.
Souza, A. S., Overkott, C., & Matyja, M. (2021). Categorical distinctiveness constrains the labeling benefit in visual working memory. *Journal of Memory and Language*, *119*, 104242. https://doi.org/10.1016/j.jml.2021.104242
Suchow, J. W., Brady, T. F., Fougnie, D., & Alvarez, G. A. (2013). Modeling visual working memory with the MemToolbox. *Journal of Vision*, *13*(10), 9. https://doi.org/10.1167/13.10.9
van den Berg, R., Awh, E., & Ma, W. J. (2014). Factorial Comparison of Working Memory Models. *Psychological Review*, *121*(1), 124–149. https://doi.org/10.1037/a0035234
van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences*, *109*(22), 8780–8785. https://doi.org/10.1073/pnas.1117465109
Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*(7192), 233–235. https://doi.org/10.1038/nature06860

Supplementary Materials

Table S1

AIC values from model comparison for aggregated data across all responses in Experiment 1

AIC	M1	M2	M3	M4	M5	M6
1st	+72	+13524	+2	+13522	**42604**	+2792
2nd	+223	+5091	+2	+5089	**50975**	+1344
3rd	+459	+446	**55081**	+448	+3	+4778
4th	+425	+118	**55094**	+128	+11	+7676
5th	+463	+204	**55013**	+249	+44	+7743
6th	+389	+285	**54943**	+304	+19	+7794

Note. The AIC values of the best-fitting iteration from each model for aggregated data in Experiment 1. A lower AIC value indicates a better-fit to the observed data. The best-fitting model for each response is bolded.

Table S2

AIC values from model comparison for aggregated data across all responses in Experiment 2

AIC	M1	M2	M3	M4	M5	M6
1st	+7	+12483	**43630**	+12549	+9	+3041
2nd	+34	+5081	**51050**	+5088	+5	+1428
3rd	+534	+569	**55081**	+591	+22	+4732
4th	+1141	+251	**55094**	+302	+48	+7739
5th	+1052	+136	**55013**	+210	+74	+8073
6th	+1084	+158	**54943**	+238	+77	+8330

Note. The AIC values of the best-fitting iteration from each model for aggregated data in Experiment 2. A lower AIC value indicates a better fit to the observed data. The best-fitting model for each response is bolded.