I'm not sure that curve means what you think it means:

Toward a [more] realistic understanding of the role of eye-movement generation in the Visual World Paradigm

ONLINE SUPPLEMENT

Bob McMurray
Dept. of Psychological and Brain Sciences
Dept. of Communication Sciences and Disorders
Dept. of Linguistics
Dept. of Otolaryngology
University of Iowa

S1. The reliability of stochastic processes: Methods

This simulation was conducted to determine the contribution of random sampling to test-retest reliability. As we describe in the main text, for any given underlying probability (say 0.5) if there are only a few trials, the likelihood of actually observing that same probability may not be high. Consequently, test-retest reliability (the likelihood of observing that probability twice ina row) should suffer.

This is likely to differ not only as a function of number of repetitions, but also the magnitude of the underlying probability itself. In some ways this should reflect basic mathematics (since the standard deviation of an observed probability is a function of that probability and the number of trials), though it is not immediately clear how to derive this function. However, estimating these values is crucial for understanding test-retest reliability of any experimental or psychometric measure – the laws of sampling set a sort of baseline expectation (that could be further moderated by the way items are combined, by variation from test to retest and so forth).

Methods. To assess this, we conducted a simple Monte Carlo simulation. We started by determining a set of probabilities of interest. These were done in logit space (which is effectively infinite, unlike probabilities), and ranged from -4 (p=.018) to 0 (p=.5) to +4 (p=.982) in steps of .5. On each run then we randomly drew 50 subjects. Each subject had a mean given by the current probability, with Gaussian variation of ±0.5 (in logit space). Thus, for simulations of and underlying p=.12 (logit=-2), subject means ranged from p=.076 to p=.18; while for simulations at an underlying p =.5 (logit = 0), subject means ranged from p=.376 to p=.623.

After setting the underlying probability for a subject, we next sampled some number of draws from that probability and computed the mean. We then did this again (for retest) and saved the results. After doing this for all 50 subjects, we computed a simple Pearson correlation between the means for test and retest.

S2. Free parameters in the model

Table S2.1 Means, SD and bounds of the random distributions for the logistic function.

Parameter	Mean	SD	Lower Bound	Upper Bound
Lower asymptote (*b*)	0.115	0.12	0	0.3
Upper asymptote (*p*)	0.885	0.12	0.5	1
Crossover (*xo*)	765	85	300	1100
Slope (*s*)	0.0016	0.0007	0.0009	0.01

Table S2.2. Means, SD and bounds of the random distributions for the asymmetric Gaussian function used to model competitor fixations.

Parameter	Mean	SD	Lower Bound	Upper Bound
Peak time (μ)	630	77	300	1300
Peak Height (*ht*)	0.18	0.05	0.05	0.35
Onset Slope (σ_1)	130	30	50	250
Offset Slope ($\sigma 2$)	250	120	50	400
Onset Asymptote (b_1)	0.05	0.015	0	0.15
Offset Asymptote (*b2*)	0.05	0.015	0	0.15

S3. A Multinomial approach (Simulation 4)

One limitation of the simulations presented in the main text was that they assumed a binomial distribution of fixations, centered around the underlying curve. That is, fixations to each object were modeled simply as whether or not the participant was fixating the target (or competitor). This is an oversimplification of the true, *multinomial* process. That is, at each time, participants are choosing which of the four objects to fixate – a choice to fixate the target necessarily means a choice not to fixate the cohort, or unrelated objects.

Supplement S3 thus sought to explore whether this additional step toward realism would constrain the key points of this study. We constructed a simple multinomial version of the models in Simulations 1-3. These were run under all three fixation models (HFS, FBS, and FBS+T), and we conducted a simplified analysis.

Approach

The multinomial model was intended to model a typical an experimental situation with a target, an onset or cohort competitor that would be expected to have heightened competition, and two unrelated objects that would be expected to show lower competition. The target was assumed to have an underlying fixation curve that took the logistic form; the cohort and unrelated objects came from an asymmetric Gaussian. The underlying parameters of these functions were random for each subject and unconstrained with three exceptions.

First, the lower asymptote of all four functions was assumed to be the same. Second, the unrelated objects had lower peaks than the cohorts. Third, we constrained the total system to have a summed probability of less than 1.0. To accomplish this, for each subject, we drew a set of 19 parameters, and then computed the underlying curves for all four objects over time. These were summed. If the sum at any point exceeded 1.0, then the parameters were redrawn. Note that the four objects could sum to less than 1.0 (and always did at early points); this captures the fact that participants often do not look at any of the four objects.

Once the underlying fixation curves for a subject were drawn, I next generated a series of fixations. This time, rather than simply saving whether the participant was fixating the object (or not), I computed which of the four objects was being fixated at each time. This same procedure was followed for all three fixation models. Note that for the FBS+T model, when fixations were directed to the cohort, the unrelated, or to nothing I used the base distribution of fixation durations (with the shorter means), if it was directed to the target, I used the longer ones.

After generating the fixation curves to each object, data were averaged and curves were fit separately to the target, cohort and unrelated items. Lastly, these were correlated with the true values, and bias was quantified as before. Our analysis focused solely on the targets and cohort. The unrelated items were very difficult to fit accurately (since fixations were so low), and would not be typically of interest in a VWP study.

Results

As Table S3.1 shows, fits to both the targets and cohorts were very good overall, in all three simulations (HFS, FBS and FBS+T). In general targets averaged well above 0.99 and cohorts

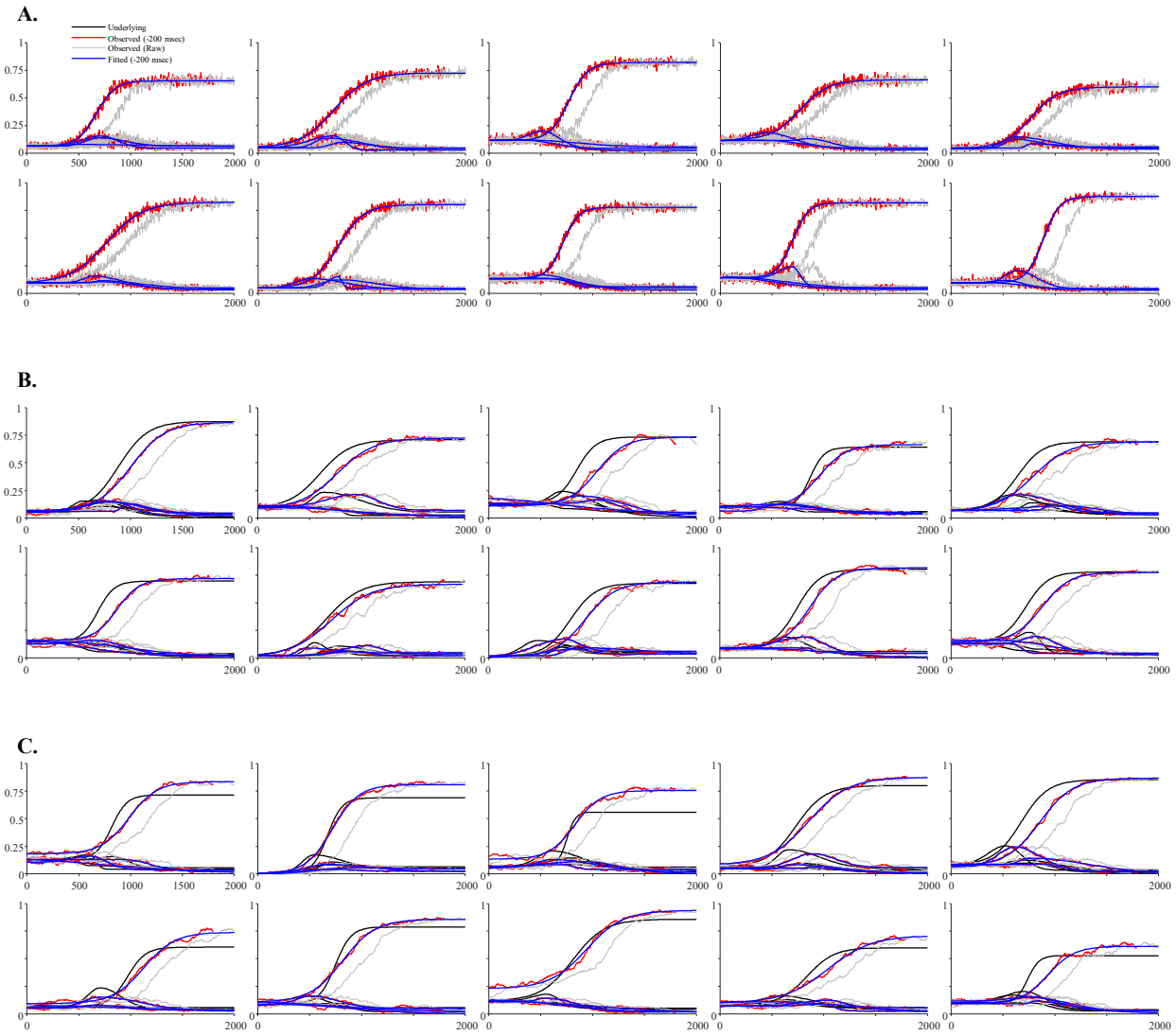

Figure S3.1. Representative subjects for the multinomial model for A) HFS; B) FBS; and C) FBS+T. In HFS, the observed fixation curves (red) generally match the underlying generating functions (in black). However, FBS and FBS+T show increasing divergence.

were above .93. Only a small number of runs were excluded from analysis (because either the target or cohort fit had a correlation below 0.8).

Figure S3.1 shows a representative set of 10 runs from each of the three simulations. Much like was seen in the binomial models, under the HFS fixation model, the underlying generating functions (in black) generally match the observed data (in red) fairly closely once the 200 msec oculomotor delay is accounted for. However, with the FBS model (Panel B) we see increasing divergence, particularly in the slope and timing of the functions. The FBS+T model shows even more divergence, but now the asymptotes are no longer accurate (Panel C).

Validity. We analyzed the validity of the observed parameters to the underlying ones using a correlational analysis. Table S3.2 shows the results for each simulation, along with the corresponding results from the binomial analysis presented in the main text.

In the HFS model, the target correlations were near perfect in the multinomial model (as

they were in the binomial). Cohort parameters generally showed similarly high correlations between the underlying and estimated values, and generally matched the binomial model. There were two exceptions to this: peak time showed a good correlation (r=.729) that was noticeably lower for the multinomial model than the binomial (r=.988); and the onset slope (r=.212) was substantially lower than in the binomial models (r=.946).

Table S3.1. Quality of fits to targets and cohorts. R is the correlation between the averaged observed data and the fitted line. Runs were excluded if either the target or cohort fits had a correlation below 0.8.

		HFS	FBS	FBS+T
R (fit quality)	**Target**	.997	.998	.998
	Cohort	.933	.961	.960
Excluded (out of 1000)		37	17	14

A similar pattern emerged for the FBS models. For targets, correlations were close between the binomial and multinomial model, but generally a bit lower. However, for cohorts, many parameters that were quite highly correlated with HFS assumptions dropped substantially – offset slope, for example, went from r=.909 to r=.306. However, these were generally in line with the performance of the binomial model and confirm what we saw before.

For the FBS+T model things showed a different pattern. For targets, correlations were lower for the two asymptotes, but slightly higher for the slope and crossover. However, for cohorts, correlations dropped across the board (relative to the FBS or the binomial model).

Thus, with respect to the validity of the observed data (relative to the underlying generating function), the adoption of the multinomial model showed a similar pattern to the binomial model, and largely made the situation worse, particularly at more realistic generating functions.

Bias. We next examined bias of the nine parameters estimates. Table S3.3 shows in terms of Cohen's D (mean / SD) for each parameter in each of the three simulations. Bias also showed a similar pattern to the binomial models. The HFS model was essentially unbiased, with the exception of the two timing parameters: the target crossover was delayed by 200 msec, and the cohort peak time was delayed by 184 msec. These reflect the built-in oculomotor delay. Much like the binomial model, the FBS model showed an even greater bias in the crossover (M=327 msec) and cohort peak time (M=307) that exceeded the mean fixation time. However, now bias was also beginning to appear in target slope (D=1.24) and height (D=.69). This was also observed in the binomial model. In the FBS+T model, all of the target parameters now showed substantial bias (D>1), as did virtually all of the competitor parameters.

Table S3.2. Correlations between the estimated and underlying target and cohort parameters for the three simulations.

		HFS		FBS		FBS+T	
	Parameter	Multi-	Bi-	Multi-	Bi-	Multi-	Bi-
Target	Baseline (*b*)	.999	1.0	.975	.987	.929	.935
	Max	1.0	1.0	.974	.986	.845	.892
	Crossover (*xo*)	1.0	1.0	.849	.855	.805	.792
	Slope (*s*)	1.0	.998	.755	.751	.702	.616
Cohort	peak time (μ)	.729	.988	.410	.502	.268	
	peak height (*ht*)	.984	.997	.823	.845	.762	
	onset slope (σ_1)	.212	.946	.070	.112	.040	
	offset slope (σ_2)	.909	.990	.306	.327	.222	
	offset asymp (b_2)	.995	.996	.694	.681	.550	

Table S3.3. Bias (Mean difference between observed and underlying value, and Cohen's D) of the estimated target and cohort parameters for the three simulations. For corresponding means in the binomial models see corresponding tables in the main text.

		HFS			FBS			FBS+T		
		Multi-		Bi-	Multi-		Bi-	Multi-		Bi-
	Parameter	M	D	D	M	D	D	M	D	D
Target	Baseline (*b*)	<.0001	-0.02	-.03	0.0033	0.34	.24	-.03	-1.00	-1.08
	Max	<.0001	-0.05	-.06	<.0001	<.0001	.09	-.082	-1.84	-1.12
	Crossover (*xo*)	-200.	-79.87	-88.68	-327.1	-6.23	-6.44	-328.9	-5.33	-4.88
	Slope (*s*)	<.0001	-0.03	<.001	<.0001	1.24	1.22	.0005	1.26	1.41
Cohort	peak time (μ)	-184.6	-2.47	-16.80	-307.3	-1.67	-2.13	-266.1	-1.23	
	peak height (*ht*)	-.0018	-0.21	-.06	0.019	.69	.85	.033	1.04	
	offset slope (σ_2)	-6.40	-0.16	.01	-61.6	-.38	-.27	-90.5	-0.53	
	offset asymp (b_2)	<.0001	0.07	.003	0.0026	.19	.14	.016	1.09	

Conclusions. In short, findings were quite similar in the multinomial formulation to the binomial models presented in the main text. I observed a fairly close alignment between the underlying and observed fixation curves under the HFS fixation model with little bias (outside of the oculomotor delay). However, under the more realistic FBS and FBS+T models I saw reduced correlations between the observed and estimated parameters (particularly for timing parameters), and greater bias. This also accords with the binomial models. One notable exception was that the cohort looks in the FBS+T model showed even more bias than they did in the FBS model. This is noteworthy, because the FBS+T fixation model does not lengthen the fixations to the competitor (only to the target); so the competitors should have behaved more like FBS. However, In the multinomial model the, the longer target fixations can affect competitor fixations since if the subject is fixating the target for longer, he or she cannot be fixating the competitor.

The multinomial model is clearly yet another step toward realism over the binomial models presented in the main text. Yet as I have consistently observed, each step toward a more realistic fixation generating function leads to lower validity estimates and increased bias of the observed data relative to the underlying.

S4. Supplementary Reliability Figures

One question that arises in simulations on reliability (Simulation 4) was whether low reliability in the FBS and FBS+T simulations was due to the quality of the fits, or if the generated fixation curves were unreliable. That is, with a fixed number of trials, it is possible that any two runs of the exact same model might yield measurably different dat. To evaluate this, we created grid plots similar to Figures 4,6,8,10,12 that show the underlying and observed data as well as the quality of the fits. Critically, here the important thing is not the fits, but the generated raw data.

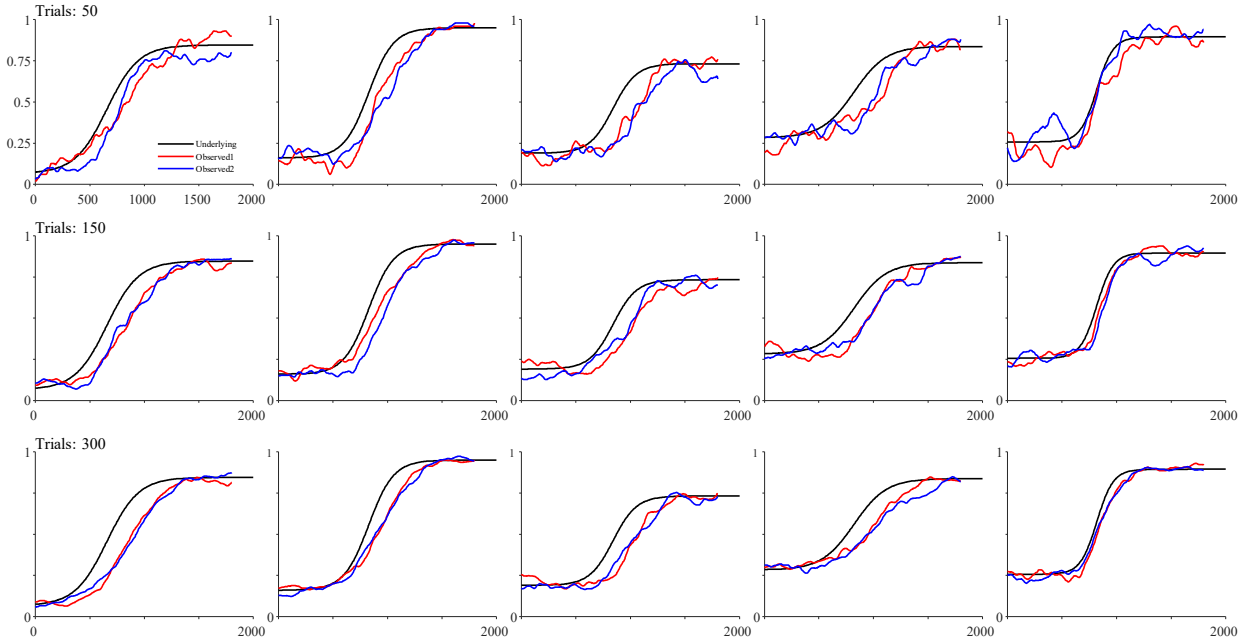

Figure S4.1. Representative subjects for target fixations in the reliability analyses assuming fixation-based sampling (FBS). Each column is a subject; rows correspond to the number of trials (top: 50; second: 150; third: 300). Black curve represents the underlying function used to generate both runs. Red and blue curves are observed fixation curves for two independent runs.

Here we show those grids for the target functions under the FBS model (Figure S1.1) and FBS+T model (Figure S1.2), and the competitors under FBS (Figure S1.3). Under the FBS assumption, the observed data (the comparison between the blue and red lines) is fairly unreliable with 50 trials (top row). However, by 150 there are fewer differences (perhaps only the middle panel) and at 300 they are quite close. Under FBS+T assumptions (Figure S1.2), the results are quite similar. However, the competitor (Figure S1.3) is a different story. Here at every level of the number of trials (including 300) there are at least a few subjects with substantive differences between test and retest. This is notable given that the observed data was generated from the same underlying function.

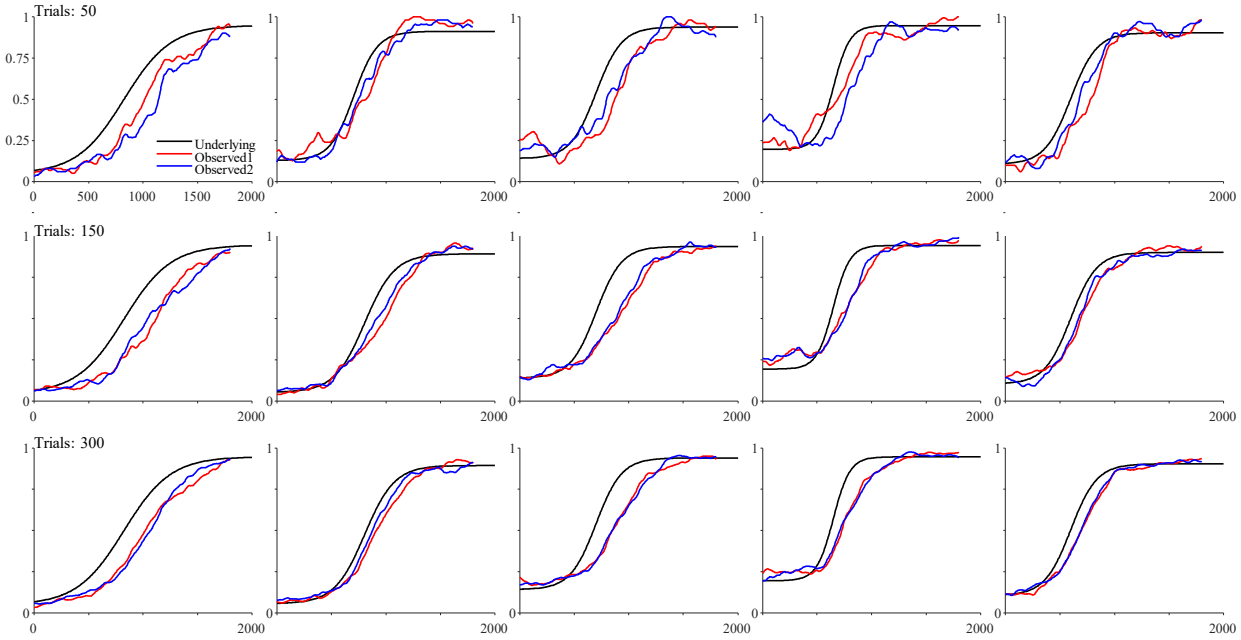

Figure S4.2. Representative subjects for target fixations in the reliability analyses assuming fixation-based sampling with enhanced target duration (FBS+T). Each column is a subject; rows correspond to the number of trials (top: 50; second: 150; third: 300). Black curve represents the underlying function used to generate both runs. Red and blue curves are observed fixation curves for two independent runs.

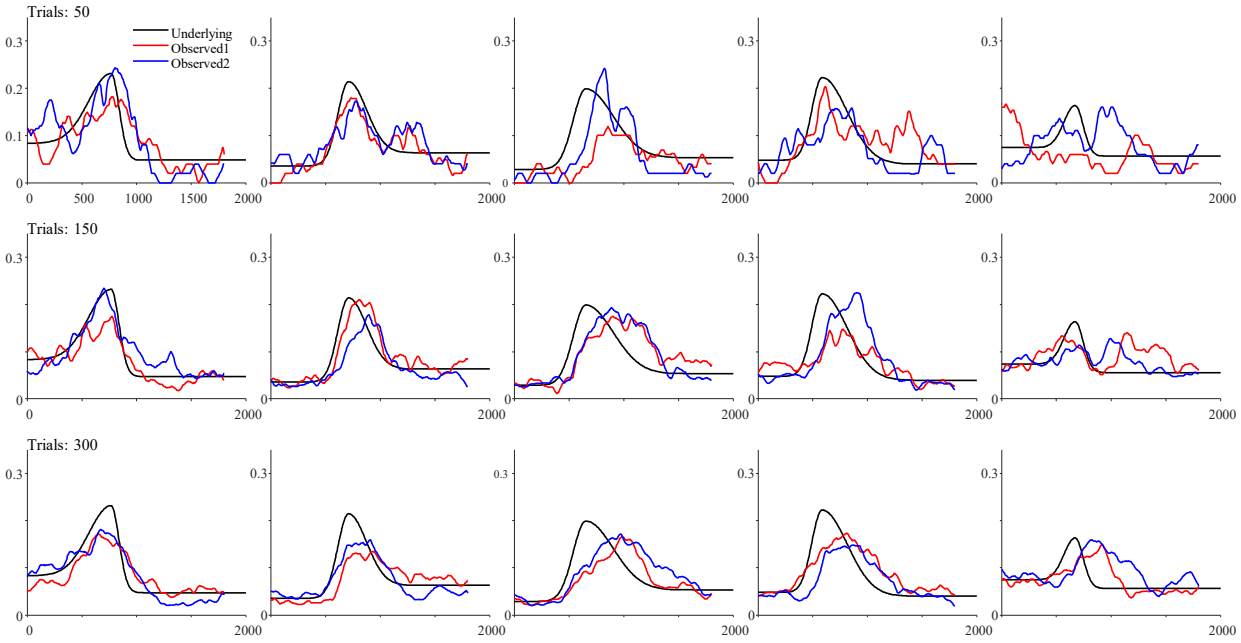

Figure S4.3. Representative subjects for competitor fixations in the reliability analyses assuming fixation-based sampling (FBS). Each column is a subject; rows correspond to the number of trials (top: 50; second: 150; third: 300). Black curve represents the underlying function used to generate both runs. Red and blue curves are observed fixation curves for two independent runs.

S5. Implementation of the Approximate Generative Model

Two issues arose with fitting the approximate generative model to the data: how to do the fitting itself, and how to estimate starting parameters.

First, standard gradient descent techniques for minimizing error require a smooth error function relating degree of error to the underlying parameters, as gradient descent relies on the derivative of that function to search the parameter space. That is, does the error go down or up (and by how much) when the slope (for example) is increased? In our case, the derivative cannot be computed analytically (e.g., derived) because the function itself is stochastic. Consequently, most gradient descent methods work by generating the error at two values of the parameter and computing the differences. However, our reliability simulations show that under a reasonable number of trials, the same underlying generating function can give rise to differences in the observed data from run to run – even when the parameters are the same. This makes the error gradient less than smooth and as a result, a curvefitter may jump around in parameter space, responding not to differences in the error, but to random variation from run to run. To solve this, the generative function creates estimated fixation curves for a large number of trials (10,000). Thus, the error function is the difference between the observed data and the data generated by the current parameters *in the limit*.

Second, even with this assumption, it was not wise to overly trust the derivative of the error function. Thus, a pattern search (Hooke & Jeeves, 1961) optimizer was used. Pattern search is somewhat coarser than derivative-based algorithms, but less sensitive to discontinuities in the data. At the broadest levels, pattern search accomplishes this by manipulating the parameters in a single direction, determining if that change improves or hurts performance, and then locking that choice (if it improved performance). It then repeats the process with the step size halved until no further changes are observed.

Third, under this scheme (and with all nonlinear approaches) one must estimate starting parameters for the underlying function. These are generally fairly straightforward for the standard curves – the *peak height* of the asymmetric gaussian, for example, can be estimated as the highest value of the data, or *crossover* of the logistic as the time where the data crosses 50%. However, under the generating models assumed here, the underlying function is not a veridical match of the observed data, making these kind of approaches impractical. I experimented with several approaches. Some parameters could be estimated directly from the data where the measures were unbiased (e.g., the asymptotes of the logistic under an FBS model). For others, I attempted a brute force search along a single dimension. Finally, I also investigated the possibility of first fitting a traditional function to the observed data, and then using a regression (based on the prior simulations) to map the parameters of the surface data to the likely parameters of the underlying data. The final procedure used a hybrid using different approaches for each parameter (see this project's OSF site for more detail about how starting parameters for both these fits and the more traditional fits are estimated: https://osf.io/wbgc7/).

S6. A lagged correlation approach to relating real time and time in the fixation curves.

While typical approaches to analysis assume a fixed lag (usually of 200 msec) between the underlying activation curves and the observed fixations (Allopenna, Magnuson, & Tanenhaus, 1998). However, some of the findings presented here suggest a more nonlinear relationship such that the delay may be greater later in the trial.

To quantify this, we conducted a lagged correlation analysis. For this analysis, a set of subjects were generated. At each timeslice, the data and the underlying probabilities were correlated to determine how the degree to which the fixation curves were related to the underlying data. We then introduce a lag, correlating the fixation data at time T to the underlying activation at T-1 and recomputed the correlation. After examining a range of possible lags, we identified the lag with the greatest correlation as the empirically derived time delay at that time. This was then done for each timeslice (of real-time). Finally this was done for targets and competitors for all of the fixation generating functions.

Approach. This approach was done as a series of 25 experiments. Each experiment ran 250 participants for 300 trials using identical procedures to the validity simulationms Lags were considered from -600 to 100 msec. The best lag at each time was computed by first computing the correlations across all lags. These were then normalized such that the lowest correlation was 0 and the highest was 1.0. Next, all of the timeslices whose normalized correlation was greater than 0.97 (97% of maximum) were extracted. The average of these was the best lag, and the SD was saved as a measure of how broad or specific the lag time was. Finally, means and SDs were averaged across all experiments to create the results presented in Figure 26.

References

Allopenna, P., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye-movements: evidence for continuous mapping models. *Journal of Memory and Language, 38*(4), 419-439.

Hooke, R., & Jeeves, T. A. (1961). "Direct Search" Solution of Numerical and Statistical Problems. *J. ACM, 8*(2), 212–229. doi:10.1145/321062.321069