What, me saccade?

Abstract

Basically there is the vwp and it is used as a proxy for word recognition. This use follows from allopenna 1996, in which he showed the the proportion of fixations to different referrents matches what would be expected with TRACE (after suitable transformation, of course). This resulted in over two decades of vwp use for such purposes. In 2021, mcmurray asked if that curve really was what we thought it was. Through an analysis of generating hypotheses of increasing (but still relatively minimal) complexity via simulation, bob showed that even in cases of moderate complexity, we were not able to positively recover the generating curve responsible for eye movements. why was this, and what are the implications for understanding these curves? In this paper, we revist the 2021 princess bride paper and offer an explanation for the demonstrated bias of the simulations. from this, we propose a new method for using vwp data to estimate underlying activation curves. finally, we top this all off by comparing the proposed method to allopena 1996 in a head-to-head, single elimination match up, a battle suitable for mount olympus itself in which two ultimate theories fight to a bloody and violent death to see which has been ordained by god to reign as total champion in the hearts and minds of language psychologists the world over.

Notes:

1. Sections that are more "narrative" are less fleshed out. This includes VWP, TRACE, general history, etc.,
2. Citations are hard coded in here awaiting a bib to be created
3. Some plots/graphics need to be redone for size
4. There is some meta commentary, partially for the reader, mostly for me
5. Spell check in my IDE is also not trivial (long story), so I have not really done that either (sorry)

1 Introduction

Spoken words create analog signals that are processed by the brain in real time. That is, as the spoken word unfolds, a cohort of possible resolutions are considered until the target word is recognized. The degree to

which a particular candidate word is recognized is known as activation. An important part of this process involves not only correctly identifying the word but also eliminating competitors. For example, we might consider a discrete unfolding of the word "elephant" as "el-e-phant". At the onset of "el", a listener may activate a cohort of potential resolutions such as "elephant", "electricity", or "elder", all of which may be considered competitors. With the subsequent "el-e", words consistent with the received signal, such as "elephant" and "electricity" remain active competitors, while incompatible words, such as "elder", are eliminated. Such is a rough description of this process, continuing until the ambiguity is resolved and a single word remains.

Our interest is in measuring the degree of activation of a target, relative to competitors. Activation, however, is not measured directly, and we instead rely on what can be observed with eyetracking data, collected in the context of the Visual World Paradigm (VWP) (Tannenhaus 1995). In the last few years, researchers have begun to rexamine some of the underlying assumptions associated with the VWP, calling into question the validity or interpretation of current methods. We present here a brief history of word recognition in the context of the VWP, along with an examination of contemporary concerns. We address some of these concerns directly, presenting an alternate method for relating eyetracking data to lexical activation. Finally, we show consistency of our method with existing continuous mapping models of activation using empirical data along with predictions made by TRACE.

This section needs work but it mostly covers the gist of what I am trying to convey, namely we are about to go from history → current state of the world → proposal and comparison → validation.

2 A brief history

We begin with a brief history to give context to later discussion. In particular, we will consider one of the leading theoretical models in speech perception, TRACE, followed by the introduction of the leading experimental paradigm, the VWP. We examine empirical evidence for the relation between these, and relevant theoretical advancements that have been made. Topics here are presented only briefly and limited to those directly relevant to the present work. For a fuller discussion of the history and uses of vwp, use google. (Or Huettig 2011b?)

An outline of the presentation (for internal use only):

1. TRACE in 1986 along with connectionist model of language
2. VWP by Tannenhaus 1995
3. VWP + TRACE, allopenna 1996
4. As far as I can tell, it's Bob's 2010 paper that was first to

(a) Look at individual differences in word recognition (not counting polynomial fits) (also relevant for the "group distribution of curves" hypothesis) and

(b) Introduce parametric forms to be fit to the data (the assumption we continue to run with), or at very least, introduce ones that are interpretable

All of the paragraphs in this section are narrative and not mission critical. Need to be fleshed out

TRACE How speech is perceived and understood has been a subject of much debate for a significant portion of psycholinguistic's history. Starting in the 1980s and persisting today, many researchers subscribe to what is known as the connectionist model of speech perception. Briefly, this model posits that speech perception is best understood as a hierarchical dynamical system in which aspects of the model are either self reinforcing or self inhibiting with feedforward and feedback mechanisms. For example, hearing the phoneme \h\ as in "hit" will "feedfoward", cognitively activating words that begin with the \h\ sound. These activated words then "feedback" to the phoneme letter, inhibiting activation for competing phonemes such as \b\ or \t\. In 1986, McClelland and Elman introduced the TRACE[1] model implementing theoretical considerations into a computer model [**?**]. Maybe useful here to discuss activation, sigmoid shape, etc.,

VWP To briefly illustrate, the VWP is an experimental design in which participants undergo a series of trials to identify a spoken word. Typically, each trial has a single target word, along with multiple competitors. The target word is spoken, and participants are asked to identify and select an image on screen associated with the spoken word. Eye movements and fixations are recorded as this process unfolds, with the location of the participants' eyes serving as proxy for which words/images are being considered.

Relating TRACE to VWP It was against simulated TRACE data that Allopenna (1998) found a tractable way of analyzing eye tracking data. By coding the period of a fixation as a 0 or 1, for each referent and taking the average of fixations towards a referent at each time point, Allopenna was able to create a "fixation proportion" curve that largely reflected the shape and competitive dynamics of word activation suggested by TRACE, both for the target object, as well as competitors. This also served to establish a simple linking hypothesis, specifically, "We made the general assumption that the probability of initiating an eye movement to fixate on a target object o at time t is a direct funtion of the probability that o is the target given the speech input and where the probability of fixating o is determined by the activation level of its lexical entry relative to the activations of other potential targets." Further of note is what this linking hypothesis does not include, namely:

1. No assumption that scanning patterns in and of themselves reveal underlying cognitive processes

[1] TRACE doesn't stand for anything – the name is a reference to "the trace", a network structure for dynamically processing things in memory

2. No assumption that the fixation location at time t necessarily reveals where attention is directed (only probabilistically related to attention)

Other assumptions included here include that language processing proceeds independent of vision (Magnuson 2019), and that visual objects are not automatically activated. Or, more succinctly, it assumes that fixation proportions over time provide an essentially direct index ofe lexical activation, whereby the probability of fixating an object increases as the likelihood that it has been referred to increases.

While other linking hypotheses have been presented (Magnuson 2019), that there is *some* link between the function of fixation proportions and activation has guided the last 25 years of VWP research.

Parametric Methods and Individual Curves While there have most certainly been advancements to the use of the VWP for speech perception and recognition (and expanded into related domains, such as sentence processing and characterizing language disorders (according to Bob)), we limit ourselves here to one in particular. In 2010, McMurray et al expanded the domain of the VWP by introducing emphasis on individual differences in participant activation curves. Two aspects of this paper are relevant here. First, although they were not the first to introduce non-linear functions to be fit to observed data, they did introduce a number of important parametric functions in use today, namely the four (or five) parameter logistic and the double-gauss (asymmetrical gauss), the primary benefit being that the parameters of these functions are interpretable, that is, they "describe readily observable properites." Second, which I suppose was also introduced by Mirman (2008) to some degree (though I have not read it yet, just pulling from Bob) is specifying individual subject curves across participants. This has been critical in that:

1. The parameters of the functions describe interpretable properties
2. This made the idea of distributions of parameters for a particular group a relevant construct

Though not stated directly (given it predates bdots by 8 years), this also served as the impetus for investigating group differences in word activation through the use of bootstrapped differences in time series (Oleson 2017) and the subsequent development of the `bdots` software in R for analyzing such differences. (A history of exploring differences in group curves can be found in (Seedorff 2018)).

This brings us to the current day, where the state of things is such that TRACE-validated VWP data is widely used to measure word recognition by collecting data on individual subjects and fitting to them non-linear parametric curves with interpretable parameters. Context in hand, we are now able to introduce some of the main characters of our story, specifically how data in the VWP is understood and used.

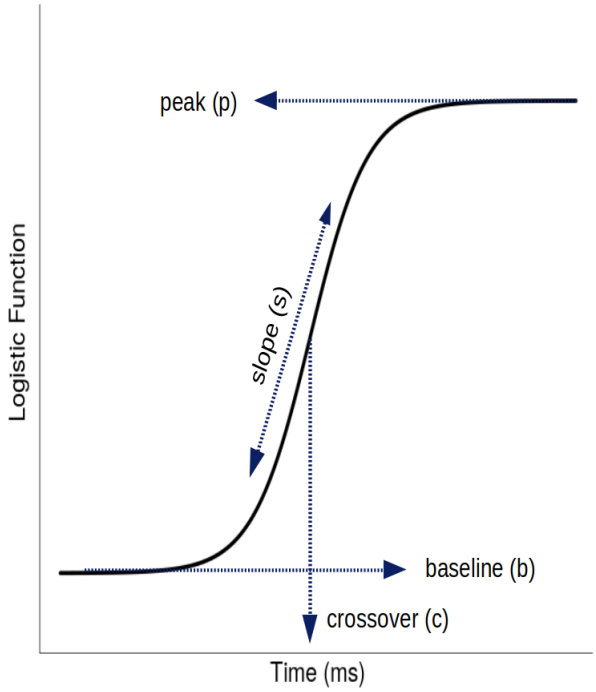


Figure 1: An illustartion of the four-parameter logistic and its associated parameters, introduced as a parametric function for fixations to target objects in McMurray 2010. Can describe the parameters in detail, but should also have the formula itself somewhere to be referenced. (Equation 1)

3 Where we are now

This section includes the finer points of the VWP, eye tracking data, and how allopena's introduction ties in with bob's parametric proposition. It may also be nice to add a section on activation specifically

3.1 anatomy of eye movements

There are three components of eye movements with which we are concerned with here. The first two, saccades and fixations, are associated with physical mechanics of eye movements; the third, oculmotor delay, is a phenomenon related to the association between cognitive activaiton and physiological response. We will briefly introduce each of these topics here.

Saccades and fixations: Rather than acting in a continuous sweeping motion as our perceived vision might suggest, our eyes themselves move about in a series of short, ballistic movements, followed by brief periods of stagnation. These, respectively, are the saccades and fixations.

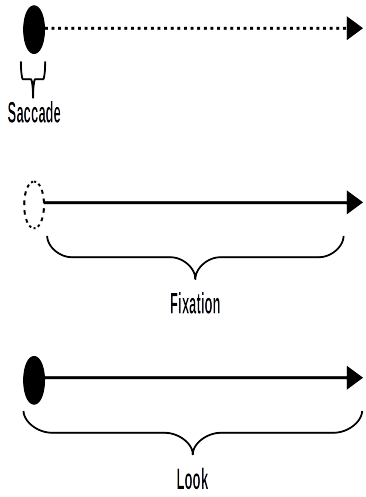


Figure 2: This image needs to be recreated for size. Illustrates saccade, fixation, and look

The short ballistic movements are known as saccades, periods of between 20ms-60ms (source? more accurate times?) in which they eye is in motion and during which time we are effectively blind. Once in motion, saccades have no ability to change their intended destination. Following the movement itself is a period of stillness known as a fixation, itself made up of a necessary refraction period from the saccade (time?) followed by a period of voluntary fixation; the typical duration of a fixation is (some length). Together, an initiating saccade and its subsequent fixation is known colloquially as a "look". See Figure 2.

Oculomotor delay: While the physiological responses are what we can measure, they are not themselves what we are interested in. Rather, we are interested in determining word activation, itself governing the cognitive mechanism fascillitating the movements in the eyes. It's suspected/stated/known (source?) that upon finishing a particular saccade, the mind is already anticipating where it will move next. Length of about 200ms also thrown around a lot. What is relevant for our purpose here, however, is that the period of oculmotor delay is a random process, resulting in biased observations between what we are able to measure and what we are interested in discovering. How this phenomenon relates to saccades and fixations is demonstrated in Figure 3.

Alternatively, there is a full figure i could use here:

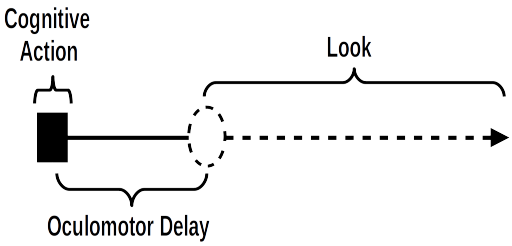


Figure 3: this also could probably be reformatted or made bigger

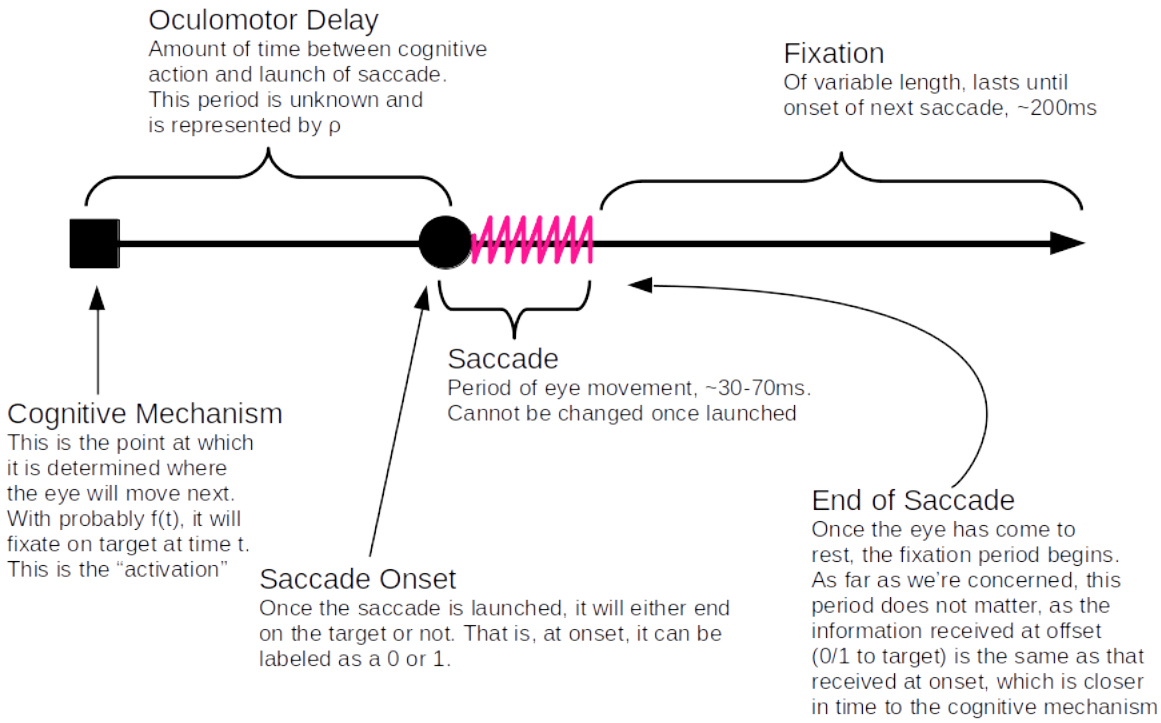


Figure 4: This figure actually doesn't look too bad, but may be better when articulating how saccades measured and why (also includes info on $f(t)$, ρ, etc., so maybe we will present this later around the time of simulation

3.2 Activation

What is it, exactly? This would be a good section to introduce notation, specifically that throughout this we will let $f(t)$ be activation in *time*, and in particular $f_\theta(t)$ where

$$f_\theta(t) = \frac{p-b}{1+\exp\left(\frac{4s}{\mathrm{p}-b}(x-t)\right)} + b. \tag{1}$$

Then I can reference Figure 1. Great, references established.

3.3 VWP data

We now consider how the aforementioned mechanics relate to the VWP. In a typical instantiation of the VWP, a participant is asked to complete a series of trials, during each of which they are presented with a number of competing images on screen (typically four). A verbal cue is given, and the participants are asked to select the image corresponding to the spoken word.

An individual trial of the VWP may be short, lasting anywhere from 1000ms to 2500ms before the correct image is selected. Prior to this, the participants eyes scan the environment, considering images as potential candidates to the spoken word. As this process unfolds, a snapshot of the eye is taken at a series of discrete steps (typically every 4ms) indicating where on the screen the participant is fixated. While there is evidence of cognition happening behind the scenes in a continuous fashion (spivey, mouse trials), an individual trial of the vwp may contain no more than four to eight total "looks" before the correct image is clicked, resulting in a paucity of data in any given trial.

To create a visual summary of this process aggregated over all of the trials, a la Allopena, a "proportion of fixations" curve is created, aggregating at each discrete timepoint the average of indicators indicating that a participant is fixated on a particular image. A resulting curve is created for each of the competing categories (target, cohort, rhyme, unrelated), creating an empirical estimate of the activation curve, $f_\theta(t)$. See Figure 5. Mathematically, it looks like this:

$$y_t = \frac{1}{J}\sum z_{jt} \tag{2}$$

where z_{ijt} is an indicator $\{0,1\}$ in trial j at time t and such that we have an empirical estimate of the activation curve,

$$f_\theta(t) \equiv y_t. \tag{3}$$

In other words, we see here that it is implicitly assumed that the trajectory of the eye follows the trajectory

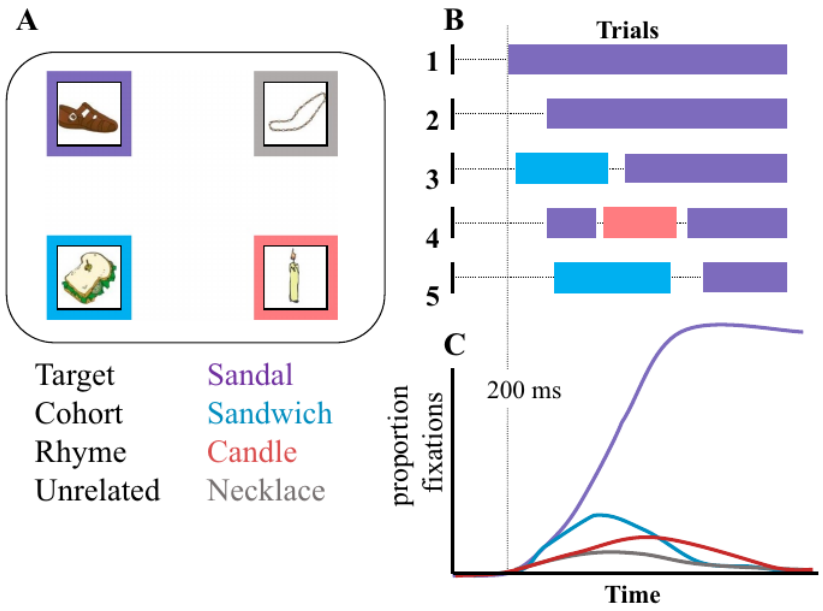


Figure 5: This screenshotted from Bob's princess bridge paper. i would like to reconstruct a similar illustration here as it does a great job illustrating the point. *However*, this section as it stands may make more sense elaborated elsewhere, in particular where I give a mathematical treatment to what the "fixation curve" is

of activation, where the average proportion of fixations at a particular time is a direct estimate of activation. As each individual trial is only made up of a few ballistic movements, the aggregation across trials allows for these otherwise discrete measurements to more closely represent a continuous curve. Curve fitting methods, such as those employed by `bdots`, are then used to construct estimates of function parameters fitted to this curve.

4 where we are going

This is a bit ahead of myself, but somewhere we need to get consistent notation and keep things in order. Here is what they are:

1. There is some idea of activation function. Under assumptions of allopena, this is basically the proprotion of fixation curve less oculomotor delay
2. There is the "generating function" which I also want to call activation. In McMurray 2022, this is the four parameter logistic that determines probability of fixating on target
3. There is the fixation proprotion curve. Until now, this has been our empirical estimate of the generating fuction, but we will show that this is not quite right. Do we give it a different function name? Can this be f_θ?
4. We will have the saccade curve. In some sense, this is similar to the proprotion of fixations in that the proportion of fixation curves is giving probability of fixating at target at some time t, whereas saccade curve is probability of a saccade launched at t landing on the target. In situation with no OM, this is an unbiased estimator or the generating curve

Ok, so now what

4.1 princess bride paper

From the abstract of this paper: "All theoretical and statistical approaches make the tacit assumption that the time course of fixations is closely related to the underlying activation in the system. However, given the serial nature of fixations and their long refractory period, it is unclear how closely the observed dynamics of the fixation curves are actually coupled to the underlying dynamics of activation."

This is a critical statement to be made, and in a sense it ties into the general idea of how we are handling the linkning hypothesis. I more or less adopt the same assumptions as Allopena, which essentially stipulates that our fixations are unpolluted index of lexical activation, independent of visual stimuli. This makes no attempt to differentiate "scanning behavior" from intendend fixations, nor do I make any assumptions regarding lenght of fixations.

n.b., bob refers to these as "fixation curve"

In 2022, McMurray brough into question the validitity of a standard VWP analysis, and a more thorough treatment of his presented arguments is warranted but for the time being counts as narrative and so the elaboration will wait. For now, we will present those elements that are crucial for understanding the direction of the methodology to be presented.

In short, the question that is being gotten at is: are we able to recover the underlying dynamics of the system in question (activation) in light of the "nature of the fixation record as a stochastic series of discrete and fairly long lasting physiologicall constrained events?" In short, the answer is no. McMurray notes that the typical, unspoken assumption implicit in VWP studies is the "high frequency sampling" (HFS) assumption, which states that the underlying activation at some time determines the probability of fixation. He then goes on to note that this is "patently" untrue and is nothing more than a polite fiction.

Nonetheless, it is useful to compare the relationship of the underlying dynamics (as we will elaborate upon further, a generating function) with the observed data in the context of the HFS, relative to other, more complex assumptions. This is done through a series of simulations, each with their own set of stochastic mechanics determining eye movements and fixations. In all cases, however, it is assumed that there is exists an underlying generating function that, at any particular time, is responsible for dictating some aspect of a subsequent fixation. We will start with an overview of the general algorithm for an individual subject, followed by a brief summary of each of the simulations.

Algorithm:

1. A set of generating parameters for the four-parameter logistic is drawn from an empirically determined distribution. This curve, $f_\theta(t)$, is treated as the probability of fixating on a target at time t
2. After a random offset start time, t_0, a binomial random event is drawn determining the probability of fixating on the target, $p \sim Bin(f_\theta(t_1))$
3. After this initial draw, a fixation occurs for a period of time:
   (a) Under the HFS assumption, this period is instantaneous – that is, whatever the time, t is also the probability of fixation
   (b) Under the FBS assumption, the length of the fixation draws from a gamma distribution, ending at time time t_2
   (c) Under the FBS+T assumption, again a random length fixation is drawn from a gamma distribution, but with a higher mean value if the fixation is drawn to the target
4. idk im desribing this weird – maybe come back to this list later

what I can do instead is offer a brief written summary of each of the methods.

High Frequencly Sampling (HFS) The underlying activation of the word *is* the probability of fixating at a particular time.

Frequency Based Sampling (FBS) The FBS assumption differs from that of the HFS assumption in that the observed data is gathered from a period of fixations of random duration. Once each fixation is "drawn", the subject remains fixated on a particular object for the full length of the fixation. The next fixation's location is determined at the *onset* of the previous fixation. In particular, this simulation assumes that immediately once a fixation is made, the subject begin's preparing to launch their next saccade

Frequency Based Sampling + Target (FBS+T) This simulation is identical to the previous with the exception that duration of fixations to the target, while still random, follow a different different distribution than fixations to non-targets, with longer durations afforded to target fixations to account for "information gathering behavior".

As the complexity of the assumptions increased, so did the observed bias in recovering the parameters of the generating function.

At any rate, I'm getting to caught up in the particulars when what I really want or need to say is quite simple. It comes down to this: *the only observed behavior governed by the generating curve is the saccade when launched.*

Saccades The entirety of the bias resulting from FBS and FBS+T are the results of two facts, or put differently, there are two sources of bias that we need to consider:

1. We were "observing" data points $\{0, 1\}$ at any time t without having observed any behavior from the generating curve at that time. Let's call this added observation bias.
2. When we did sample directly from the curve at fixation onset, we were actually sampling from the onset of the previous fixation. We will call this delay bias.

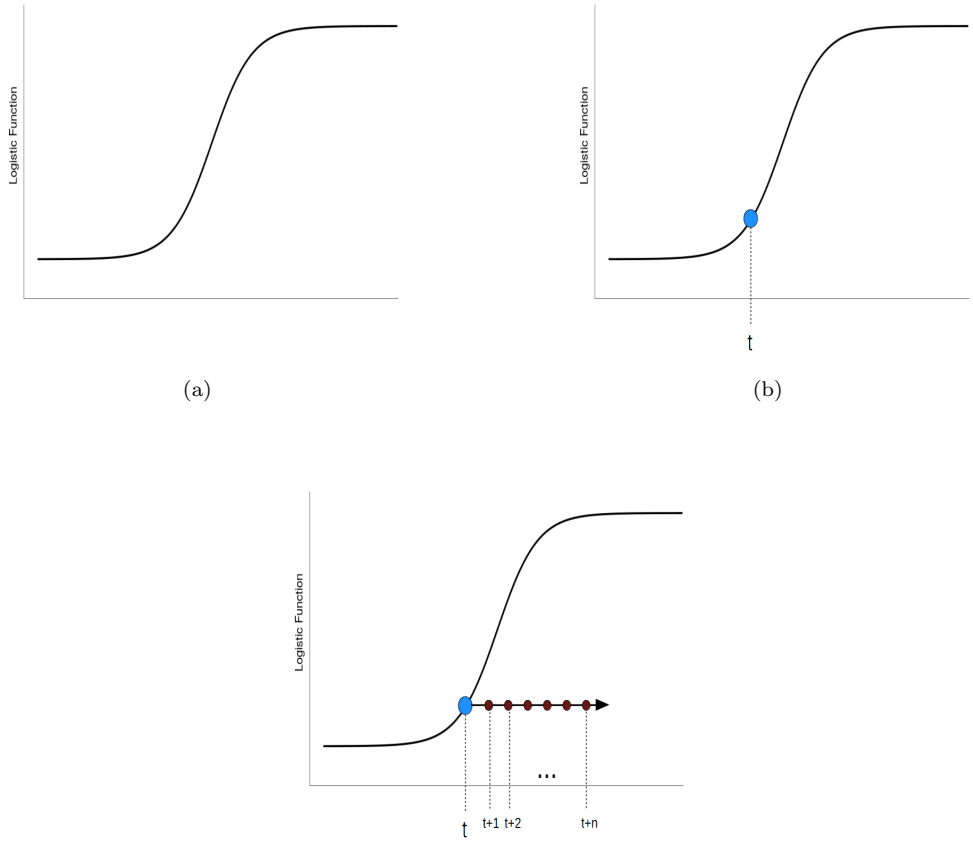
One partial solution to this, then, is to simply use the saccade data, or only collect as $\{0, 1\}$ the instance at which a fixation occurs, discarding the rest. While this does not address the delay bias, it does remove a significant amount of bias from the added observations. One obvious shortcoming is that is dismisses all "information gathering behavior" that could otherwise be gleaned from the duration of fixations. To what effect this or other enhancements may have on the efficiency of this data are yet to be seen, but at very least it offers a more clearly defensible relationship between the observed data and the generating function.

The idea of information gathering behavior is a useful one, but it assumes a linear relationship between the length of time of the fixation and strength of activation. However, one might suspect that after a period of necessary refraction, each subsequent period of time gives exponentially more weight to the argument of activation. A potential consequence of this is that an indication of fixation 50ms following the launch of a saccade may convery different information than the indication of a fixation still present 300ms following a saccade, despite the fact that these are recorded equally as $\{0, 1\}$. That is, under the present system, rather than indicating the gathering of more information, longer fixations simply increase both the bias due to added observation *and* the amount of bias on account of delay (as the subsequent fixation will have been determined further removed from its occurence when following a longer fixation). On the other hand, a mechanism for recording information-gathering behavior may be more readily implemented in a saccade-style method whereby each saccade is weighted by the length of its subsequent activation, for example.

Note too that this is consistent with a linking function in which lexical processing runs independently of visual display and is directly related to audio stimuli. It is further interesting to observe that some of the more complex mechanics (i.e., fixations to target object lasting longer) seek to introduce mechanisms by which visual stimuli contribute to word processing. In this case, however, fixations to target only impact observable behavior rather than "accelerate" the underlying activation. Given the added observation bias, this inadvertently has the effect of shifting the crossover parameter forward while also decreasing the slope.

From here, we will describe the proposed saccade method in more detail, compare the results of using saccade data against what was found in McMurray 2022, see that it largely resolves a portion of the bias, and then ask the natural question: how does this stack up against what allopenna found?

Also note somewhere that we are primarily limiting discussion to the logistic here. A brief treatise on the asymetric gaussian is given in the appendix



(a) (b) (c)

Figure 6: These illustrations can all be made larger (they were made for slides in an image editing program), but they illustrate the main point. **(a.)** here we see an example of a generating logistic function **(b.)** at some time, t, a saccade is launched (in the algorithm, a binomial is drawn with probability $Bin(f_\theta(t))$ **(c.)** at subsequent times, $t+1, \ldots, t+n$, we are recording "observed" data, adding to the proportion of fixations at each time but without having gathered any additional observed data at $f_\theta(t+1), \ldots, f_\theta(t+n)$, thus inflating (or in the case of a monotonically increasing function like the logsitic, deflating) the true probability.

4.2 Saccade method

If we are to consider eyetracking data samples from some probabalistic curve, it becomes necessary to differentiate between the two types. A saccade launched at some time, t, can be considered a sample from a data-generating mechanism at t. The duration of time between a given saccade and the one following follows a different mechanism altogether. By clearly delineating the mechanism from which we are sampling, we are able to reduce observed bias in the reconstruction of the activation curve.

In light of this, and in contrast to the fixation method, we propose estimating the activation curve with the saccade data alone. The primary benefit of this is two-fold. First, as suggested above, by decoupling two different types of data we are able to be more precise in what it is we are sampling. Second, as explained in the previous section, we are removing the added observation bias.

An important difference between these two methods is in the structure of the data itself. Whereas the former collects an array of data, with an observation for each time point in each trial, the saccade method is sparse, with the observed data indicating the outcome of the saccade, as well as the time observed. It is best represented as a set of ordered pairs, $\mathcal{S} = \{(s_j, t_j)\}$, with j indexing each of the observed saccades, and with

$$s_j \sim Bern(f_\theta(t_j)). \tag{4}$$

A value of $s_j = 1$ indicates a saccade resulting in a fixation on the target.

As with the proportion method, the observed data can be used as input for `bdots` to construct estimates of generating parameters.

5 Simulations Against Princess Bride

Here we now replicate the results of the princess bride paper, though with a few adjustments in light of the previous discussion. In particular, we noted that the two types of bias presented in the original simulations were added observation bias and delay bias. The first, added observation, we are addressing with the propsed saccade method. As to the second, we first elaborate here with a brief discussion of the delay bias and how it may also related to oculomotor delay.

In the original princess bride paper, FBS and FBS+T were only differentiated by the amount of additional bias introduced in FBS+T. Specifically, by drawing fixations to the target from a gamma distribution with a larger mean, we were both increasing the amount of added observation and delay bias, both consequence of the longer fixation period. Understanding that these differ in degree rather than kind, we collapse them

into a single construct here, as we will elaborate on shortly.

We also make adjustments to how we deal with oculomotor delay. As was shown in the simulations with the HFS assumption, a fixed oculomotor delay simply resulted in a horizontal shift of the estimated function, having otherwise no impact on the *shape* of the function. In contrast, added observation and delay bias both drastically impact the final shape. Further, in typical instances in which we are using the VWP, we are more frequently concerned with the relative difference between two curves rather than the curves themselves. The magnitude and relative location of such differences will be preserved under a horizontal shift, having ultimately little consequence on the resulting analysis.

In light of this, we will seek to combine the functional impact of oculomotor delay with the impact of more complex eye behavior in the following way: we recognize that with the exception of the HFS assumption (which is not considered here), any observation at time t will have been prompted at some time previously (that is, drawn from the activation curve). The length of this delay will be denoted $\rho(t)$. In the case of the HFS assumption, for example, this simply would have been $\rho(t) = 200ms$. For the FBS and FBS+T case, it would have been $\rho(t) = 200ms +$ length of previous fixation. As such, we can reduce the conditions under which we compare the fixation and saccade methods to two scenarios:

1. $\rho(t)$ is a constant function, including zero
2. $\rho(t)$ is a random variable, independent of the value of t_j

As such, we will not be including as a part of this the original 200ms oculomotor delay. Instead, we will consider cases in which $\rho(t) = 0$ and $\rho(t)$ follows a gamma distribution, independent of time and of current or previous fixations (I should report mean/variance).

—

Not sure yet if this notation (below) comes up

—

As in the princess bride paper, we will let $f_\theta(t)$ be a four parameter logistic, representing of generating or activation curve. Understanding that what we observe at time t was drawn from this function at time $t - \rho(t)$, we will differentiate the underlying activation curve the the observed data,

$$g_\theta(t) = f_\theta(t - \rho(t)) \tag{5}$$

Each simulation will be conducted with $N = 300$ trials, sampled from the same data generating function for each, with the attempted recovery of the generating curve done using the `bdots` package.

Note somewhere: this and all other code used available on my github

5.1 Overview

Simluations here we conducted with only mechanisms related to fixating on the target object, that is, constructed from the four-parameter logistic curve which we here call the generating curve:

$$f_\theta(t) = \frac{p-b}{1+\exp\left(\frac{4s}{p-b}(x-t)\right)} + b. \tag{6}$$

An empirical distribution of these parameters was generated from prior studies (Farris-Trimble, McMurray 2013). From this distribution, each subject drew a set of parameters to construct their own individual generating curve.

For an individual subject, a single trial consists of selecting a random onset time t_0 and, based on their generating curve, selecting with binomial probability $f_\theta(t_0)$ if a particular saccade would be directed towards the target. Following each draw, a fixation length is drawn from a gamma distribution, the end of which designates the subsequent time, t_1. Here, a second draw from the binomial is performed; the probability of fixating on the target differs depending on the method and will be explained in more detail in the following section. This process of launching saccades with some probability based on $f_\theta(t)$ with durations of fixations following a gamma distribution is repeated until the sum of all fixations in a single trial exceeds 2000. Two measures are recorded each trial: the first, a `data.frame` indicating the onset time of each saccade, along with an indicator of whether or not it was launched towards the target. Second, a fixed length vector recorded at intervals of 4ms an indicator of the fixation. This data made up the saccade and fixation data, respectively.

For gamma – empirically determined (though that doesn't really matter here), shape parameter 4.88, scale parameter 35.035. This gives us a gamma distribution with a mean value of 171.18 and a standard deviation of 77.443)

Each subject was simulated to have 300 trials, and 1000 subjects were randomly generated. For each subject, saccade data was concatenated as to preserve each saccade launch, its onset time, and whether or not it was launched towards the target object. Fixation data was averaged across trials at each 4ms period to create a proportion of fixation curve.

All saccade and fixation data was then fit to the four-parameter logistic function with the R package `bdots` (version 1.2), using the `logistic()` function with starting parameters `params = c(mini = 0, peak = 1, slope = 0.002, cross = 750)`. This was to ensure consistency in the fitting algorithm performed by `gnls` which is sensitive to starting parameters. For curves fit to fixation data, fitted functions with $R^2 < 0.8$ were not included; for saccade data, fits were excluded if the peak parameter estimate exceeded the base

parameter or if the slope or crossover were negative. Only subjects whose curves passed both criteria were included in the final analysis. In all, 996 of the original 1000 subjects were kept.

Finally, for each simulation we investigate the distribution of parameter biases between those used for the generating curves and those recovered by `bdots`. We also consider a representative collection of fitted curves for both saccade and fixation methods against the generating curve. Finally, we consider metrics of mean integrated squared error (MISE) and R^2 between the two methods under both simulation conditions.

5.1.1 Fixed Delay

Horizontal shifts in the data (such as those introduced with a 200ms oculomotor delay) do not change the shape of the data, and as was demonstrated in McMurray 2022, a simple shift under the HFS assumption was able to fully recover the generating parameters. We then choose to begin with the assumption $\rho(t) = 0$, or with no delay between when a saccade is launched and its destination. In terms detailed earlier, this means that when a fixation ends at t_j and a subsequent saccade is launched, it has a probability of fixating on the target of $f_\theta(t_j)$. As such, we should expect the saccade method to be an unbiased estimate of the generating function (and this is indeed the case as evidenced by the figures). Accordingly, the degree to which bias is introduced in the fixation method will be solely the consequence of the added observation bias.

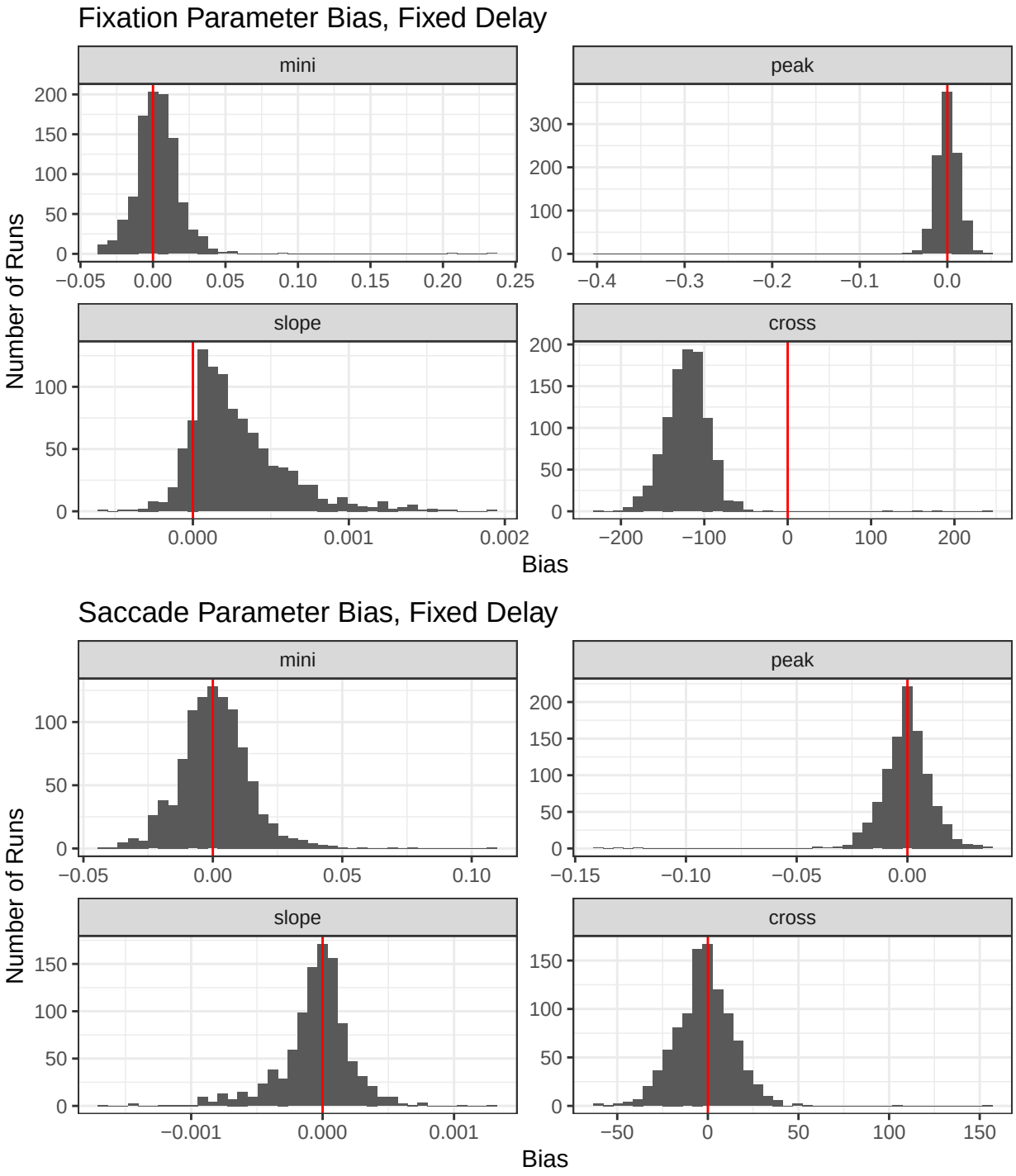


Figure 7: Distribution of parameter bias for fixation and saccade methods under fixed-delay simulation. The bias induced in the fixation method is all a consequence of the added observation bias. Somewhere note these are TRUE PARS - FIT PARS. We see evidence that added observation bias has the effect of "pulling" the curve at both ends, resulting in later crossover and less steep curves

Indeed, this is what we see when looking at the histograms of the observed bias, where negative values indicate that the fitted parameters were larger than those generating. This has the additional benefit of making theoretical sense: if the crossover point represents the time in which the probability of fixating on the target, added observations occuring *before* the crossover point would artificially inflate the number of "0"s observed, pushing the estimated crossover point forward. Further, fixations *after* the crossover point would artificially inflate the number of observed "1"s. Given the variability of the binomial is smallest with probabilities close to 0 or 1, this effect would have the largest consequences near the beginning and end of the simulated trial data, necessitating a more gradual slope in the center, which is also evidenced in Figure 7. The impact that this has on the curves themselves is also illustrated in Figure 8.

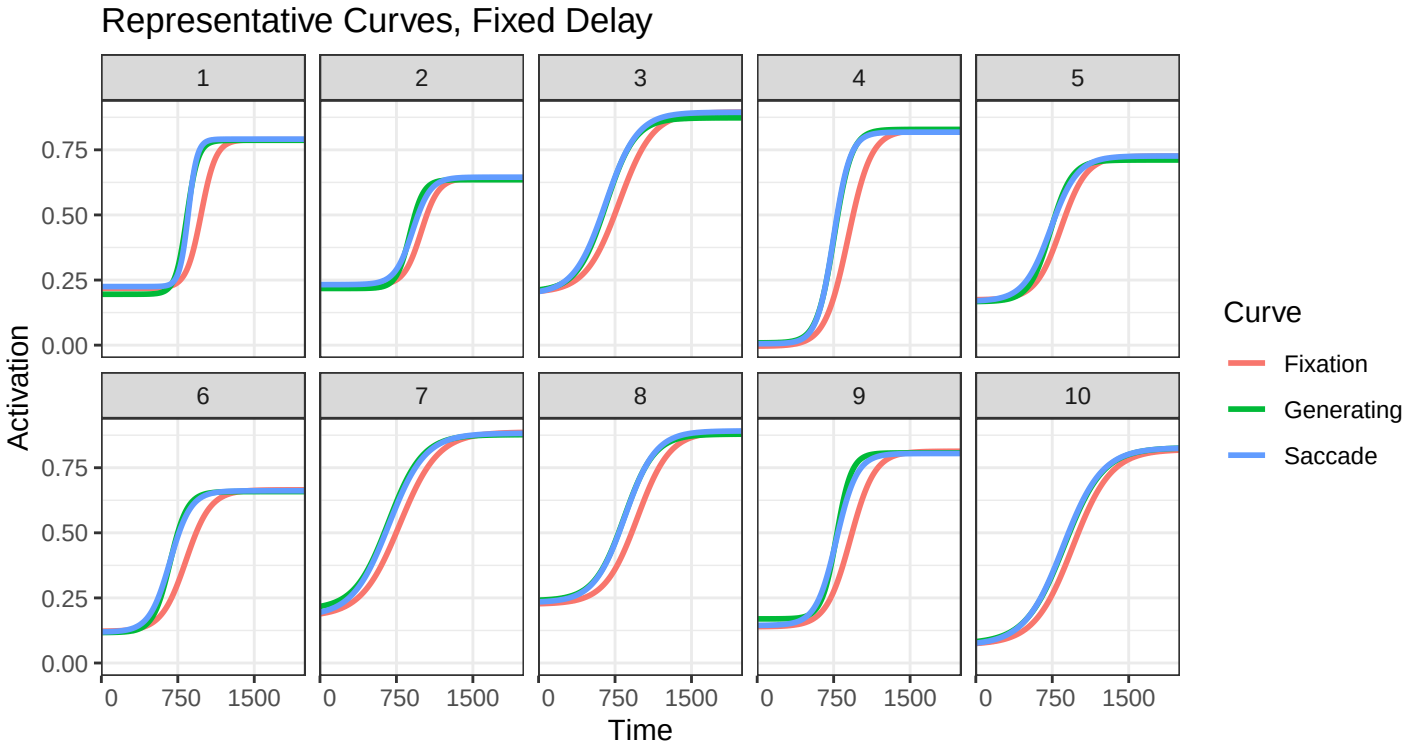


Figure 8: Representative collection of fixed-delay curve, including the generating function, as well as estimated curves from fitting data using fixation and saccade methods

5.1.2 Random Delay

While estimating the generating functions in the absence of any delay is ideal for demonstrating the added observation bias, it is a poor reflection of any actual mechanics governing the link between lexical access and physiological behavior. In the random delay simulation presented here, we modify the behavior of the simulating function in one way. Rather than assuming that a saccade launched at time t_j draws from a

binomial with probability fixating on the target at t_j, we instead assume that this probability is determined at the onset of the previous fixation. This has the effect of adding a *delayed observation bias*, with the delay following a gamma distribution with $\mu = 171.18$ and $\sigma = 77.443$. The results of this can be seen in Figure 9.

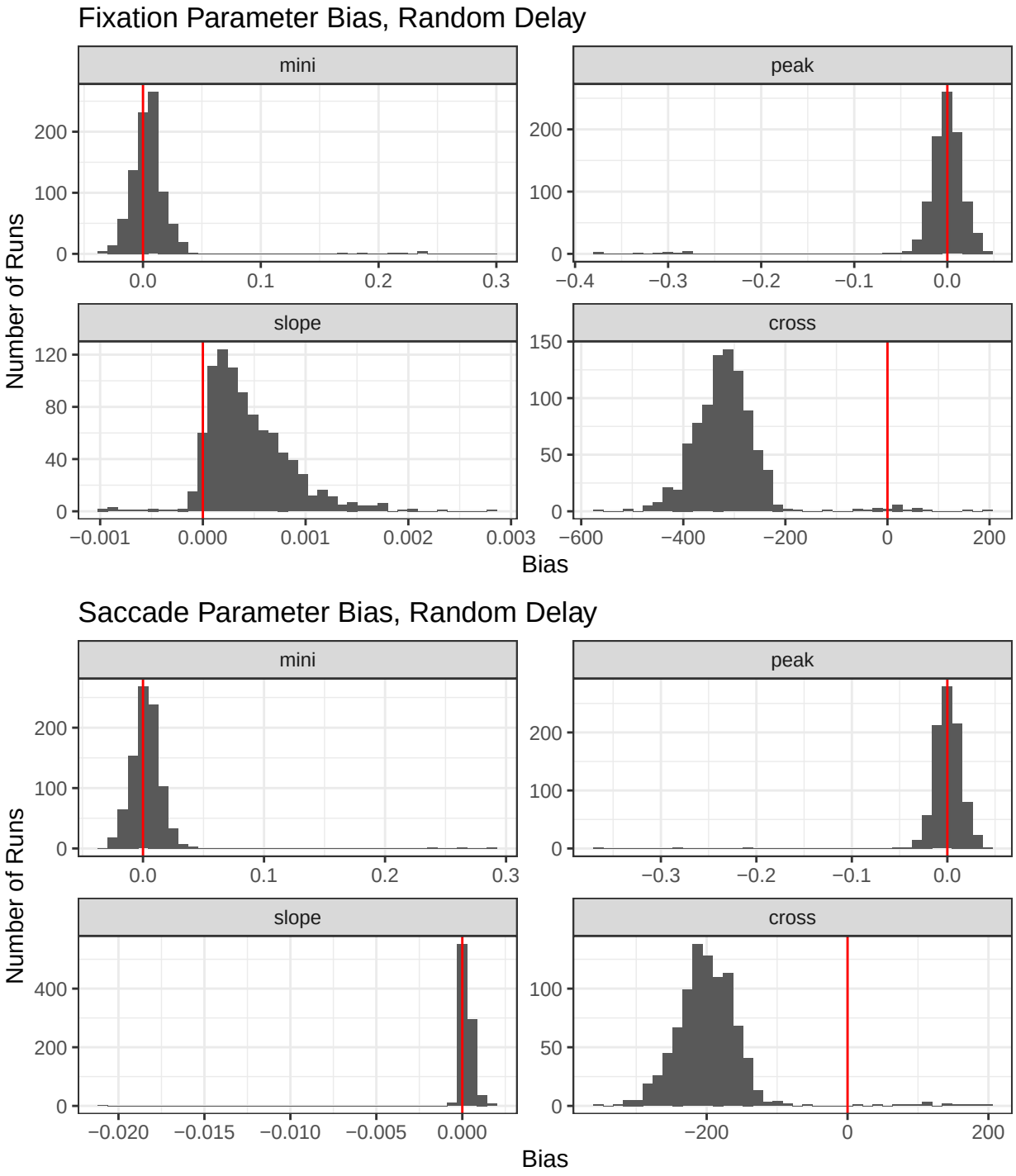


Figure 9: Distribution of parameter bias for fixation and saccade methods under random-delay simulation. The bias induced in the fixation method is all a consequence of the added observation bias AND delay bias, which has consequence of further shifting crossover parameter underestimating slope, but now with saccade too. Somewhere note these are TRUE PARS - FIT PARS. We see evidence that added observation bias has the effect of "pulling" the curve forward, resulting in later cross over and less steep curves

The behavior of the bias is quite similar to that in the fixed delay method, and the theoretical reasons governing this behavior are also similar. The exception here is that *all* observations are artifically deflated from the true probability as the generating function is monotone and any saccade launched at time t was in fact determined at time $t - \rho(t)$ where $f_\theta(t - \rho(t)) < f_\theta(t)$ for all t.

For the fixation method, we still see evidence of the "pulling" behavior at both ends of the curve, resulting in a consistent underestimate of the slope parameter; the bias of the crossover parameter is more pronounced and with greater variability, a combination of the delay observation acting in concert with the added observiation bias. Alternatively, consideration of the saccade method shows some degree of bias in the estimate of the slope parameter, with clear evidence of the delay bias in the crossover parameter. While the reasons for the bias in the crossover parameter are evident, those for the bias in the slope parameter are less so and may simply be a consequence of the lack of independence between these two when fit with `gnls`.

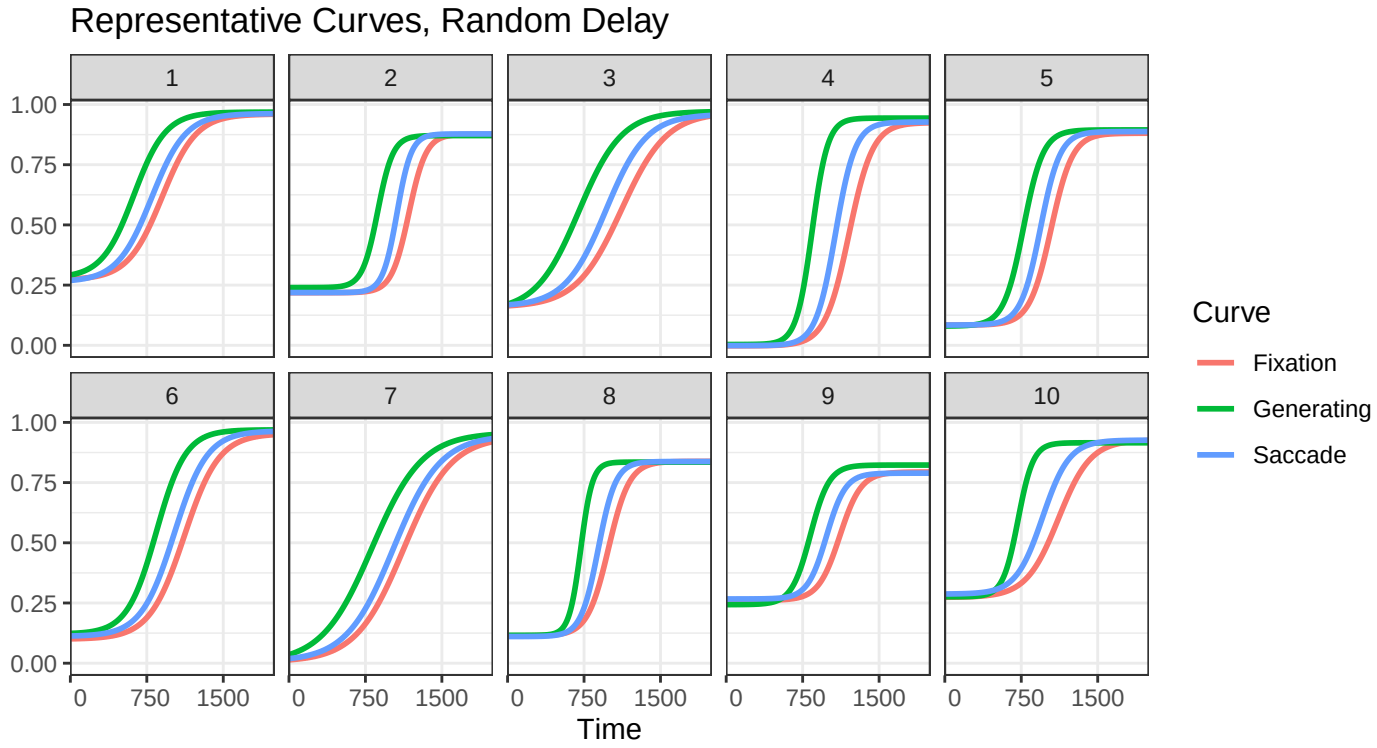


Figure 10: Representative collection of random-delay curve, including the generating function, as well as estimated curves from fitting data using fixation and saccade methods

5.2 Discussion

Perhaps unsurprisingly, Table 1 demonstrates that (1) Situations in which there is no delay between the generating function and observed behavior are easier to recover parameters and (2) the saccade method

performed much better in all these cases. This table only includes MISE, I need to add R^2, though the results will functionally be the same (to the degree maybe don't need R^2).

Curve	Delay	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Fixation	Fixed	1.95	8.18	11.40	13.28	15.98	215.67
Saccade	Fixed	0.01	0.16	0.32	0.52	0.56	78.22
Fixation	Random	20.25	50.95	68.60	73.08	90.92	192.56
Saccade	Random	5.74	21.42	29.29	33.40	40.63	185.79

Table 1: Summary of mean integrated squared error of the fits with their generating curves

As an aside, this section needs to be where I present my strongest argument, as I have come to conclude that the section on TRACE doesn't offer stronger evidence one way or the other for fixation or saccade method (I will elaborate further in that section). Parts of this may belong in general discussion at the end instead of here.

There has recently been what appears to be a renewed interest in establishing a more concrete link between lexical activation/word recognition and what is being measured with the VWP. On a grander scale, this comes down to a collection of competing hypotheses presented by Magnuson (2019). Yet, seemingly independent of the theoretical argument is what appears to be a more practical issue, namely dealing with VWP data in a tractable way. McMurray 2022 seemed to suggest that many of us carry an implicit HFS assumption and that any deviations from this (of which there certainly are) raises a number of issues in terms of what we are actually measuring with proportions of fixations in the VWP.

McMurray 2022 did not set out to establish a competing hypothesis linking theory to data; rather, he demonstrated that even under moderate deviations from HFS, what we recover is a fundamentally biased index of the mechanism we are purporting to measure. The simulations presented did commit to few additional assumptions, most generally that a generating function assume some parametric form and that this function directly determines the probability of launching a saccade (and consequently fixating) on the target object. Distilled from the particulars of these assumptions, however, are two phenomena worthy of consideration for any hypothesis linking activation to eye mechanics:

1. There is certainly a delay between the mechanism of interest and any physiological behavior, and this delay almost certainly has a random component (as opposed to a fixed 200ms delay). This includes aspects of oculomotor delay, saccade refractionary periods, etc.,. We can be agnostic to any of the particular details of this delay so long as an appropriate distribution of the delay can be estimated
2. The mechanisms governing saccades and fixations are perhaps necessarily different, and there is utility is treating them as observations from separate processes. This allows accomodation of a wider range of possibilities, from scanning behavior, information gathering from the visual world, and account for length of fixation, etc.,.

Note: relating to the first point, while it was not shown here explicly (as we only included the assumptions with FBS rather than FBS+T, there is likely a nonlinear effect to the added observation bias in light of

greater and greater delayed observation bias (that is, more added observations that themselves are more biased). We sought merely to demonstrate the existence of these biases rather than the effect, though it is likely that with investigation we could show further benefit to the saccade method when the fixation durations are more random, further distorting the "shape" of the estiamted function.

In short, what we have hoped to accomplish here is not to drastically change the original assumptions presented in Allopenna (1996) and elaborated upon in Magnuson (2019), but rather to qualify them in statistically sound ways. This starts by separating the processes governing saccade movement and duration of fixation.

As a not really conclusion, I am left to wonder to what degree the proportion of fixation method was a "local minimum" is the pursuit of utilizing eye-tracking data. The proportion of fixations created an ostensible curve, prompting McMurray to establish theoretically grounded non-linear functions to model them. These, in turn, where shown to be suitable functions with which to model saccade data over a period of trials. Had saccades leant themselves so naturally to visualizing as the proportion of fixations, perhaps that is where we may have started.

copy and pasted from elsewhere:

[this really belongs with a compare-and-contrast section following saccades. primary benefit of this....over what?] One of the primary benefits of this method is that it captures the duration of fixations, with longer times being associated with stronger activations. This becomes important when differentiating fixations associated with searching patterns (i.e., what images exist on screen?) against those associated with consideration (is this the image I've just heard?). A shortcoming, however, is that it conflates two distinct types of data, generated via different mechanisms, the fixation and saccade.

6 Compare with TRACE

I have a few issues with this section, and as I have fleshed out my reading and understanding of things, my intention with this section has changed. Originally, my hope was to show that non-linear functions fit with empirical saccade data would be a better match to what is predicted by TRACE than what is found using fixation data. This, however, seems to be the wrong thing to do. There is apparently a magnificent number of ways with which to transform TRACE activation data to probabilities of fixation, Neverminding the fact that the saccade curve is a fundamentally *different* concept/mechanism than the proportion of fixations, calling into question the value of a direct comparison as well as the validity of suggested transformations for evaluating the saccade method, including those in McMurray (2010) (though I'm not sure that this is really much of an issue, as none of these transformations had mechanics uniquely specific to the properties

of fixations).

This general idea is related to an observation made earlier by Allopenna and friends,

"It is important to note that although the TRACE simulations provided good fits to the behavioral data, the results should be taken as evidence in support of the class of continuous mappping models, rather than support for particular architectural assumptions made by TRACE."

Of course, this finds us in a bit of a circular loop – Allopenna suggested that the consistency of evidence was used to support the assumptions of TRACE, and here we are suggesting TRACE as evidence for the saccade method. What seems more appropriate, then, is to demonstrate that there *exist* transformations of TRACE data in which *either* the saccade or fixation methods creates a better fit (as measured by R^2 or MISE). As such, it makes less sense to use TRACE to show which is "better", but rather to use TRACE to demonstrate that what is estimated with the saccade method continues to be consistent with the continuous mapping model of lexical activation. Having established theoretical consistency, my argument for the saccade method will rest on what was presented in the previous section, namely the separation of saccades and fixations, and the problems illustrated with the added observation bias.

As it is, I have a few sections here addressing high level concerns. How it should be precisely organized is up for debate, but the work itself should be largely finished.

Finally, I will note here without much other detail – the entirety of this next section rests with the empirical and simulated data from McMurray 2010. For the empirical data, only N/TD subjects were used and for TRACE, only the 14 simulations with the default hyperparameters from jTRACE. The specifics of cleaning (less what I mention in relevant sections here) can be cast to the appendices

6.1 On fitting saccade data

I will go into more detail later on the precise transformations that I did to arrive at the empirical data from the raw data. One key thing to note, however, is that rather than using the separate saccade data in the Access DB, I finished cleaning the fixation data and then transformed this to saccade data based on the start time of subsequent fixations. This helped address ambiguities that resulted from deciding which saccades to be included. For example, if we neglected to include any saccades that began before the onset of the audio signal, the first saccades recorded (generally) had a probability of fixating on the target of about 0.25, resulting in an empirical curve with a base parameter much closer to 0.2. This was addressed by artificially setting the first saccade to have occured at $t = 0$, which nearly always had an associated recorded fixation not at the target. After making this adjustment, the baseline of the saccade curve matched nearly that of the fixation curve, making the saccade curve more closely match the shape of the fixation curve.

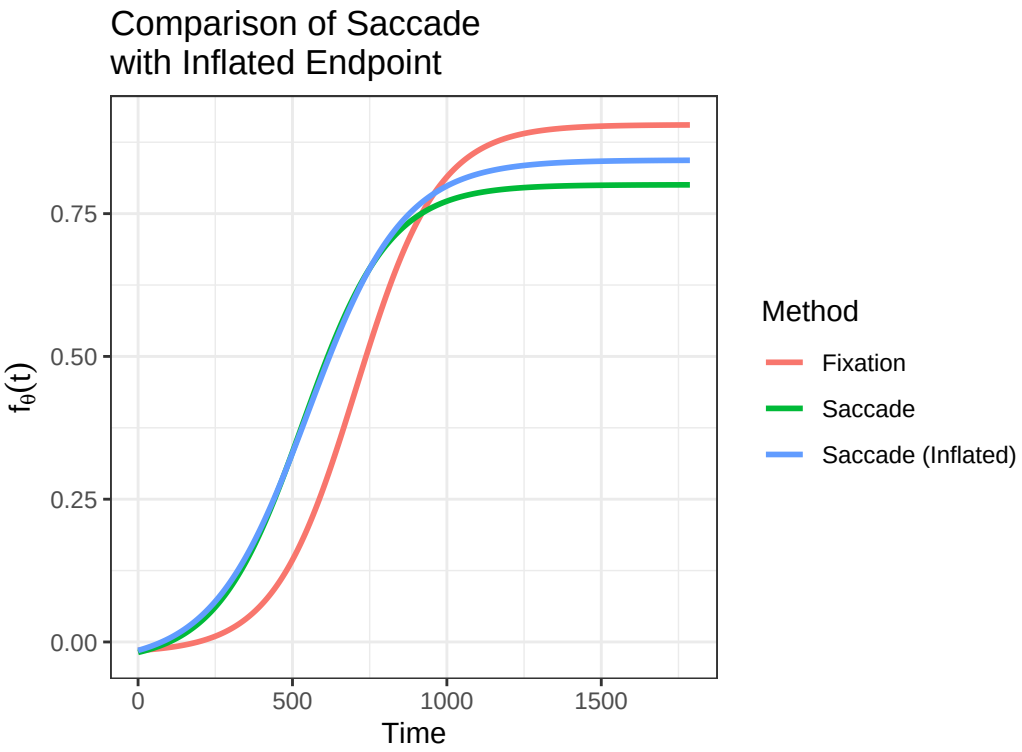


Figure 11: this is what happens when i inflate saccade with additional saccade at last endpoint

Slightly more of an indulgence was how to treat the end of the saccade curves at each trial. Necessarily in all cases, following the last saccade, recorded fixations were constant until the end of the trial, drastically increasing the "added observation" bias and resulting in a fixation curve with a peak much closer to 1.

On one hand, this could perhaps have been dealt with by addressing response times and making the appropriate adjustments. Or, far more simply (and with fewer researcher degrees of freedom), I simply added one last saccade to each trial with it's target location being that of the last fixation (typically the VWP Target). The rest of the analysis does not depend on this decision in any fashion, and as the results are functionally the same, I elected to use the saccade curve with the inflated data. This most closely matches our expectation of the relation between the fixation and saccade curves and addresses (to some degree) the asymptotic behavior of the saccades which would otherwise be uncollected. A demonstration of these differences is given in Figure 11.

6.2 Transforming TRACE data

My primary concern with the TRACE data is I am seemingly unable to reconcile it visually with what is presented in the 2010 SLI paper. Specifically, I never achieve a baseline near 0. I tried manipulating the temperature of the luce choice rule (LCR) with both constant factors and sigmoidal shapes with differening

parameters; I also tried playing with some of the parameters from the scaling factor function.

Referring back to an email we exchanged 12/14/2022, you (you being theBob) gave me a list of adjustments to make to the scaling factor, including swaping the activation and crossover, as well as expanding the exponential term to include the entire denomenator. I did this and confirmed that, as you had, the function goes from 0.0002 at maxact=-0.2 to .739 at maxact = .55. The issue, though, is that this is performed *after* luce choice rule implemented. In that situation, the minimum activation observed is 0.25 rather than -0.2. This made me think that perhaps some other permutation of transformations would result in a curve starting closer to 0 and peaking nearer to .75 (for example, scaling the raw TRACE activations). In my collection of attempts, I never found anythning to quite correct for this. It may be a bit much, but I have included plots of the TRACE data related to the target at different points in the transformation process to see how it changes. Maybe something in that will ring at bell. These are included in Figure 6.2.

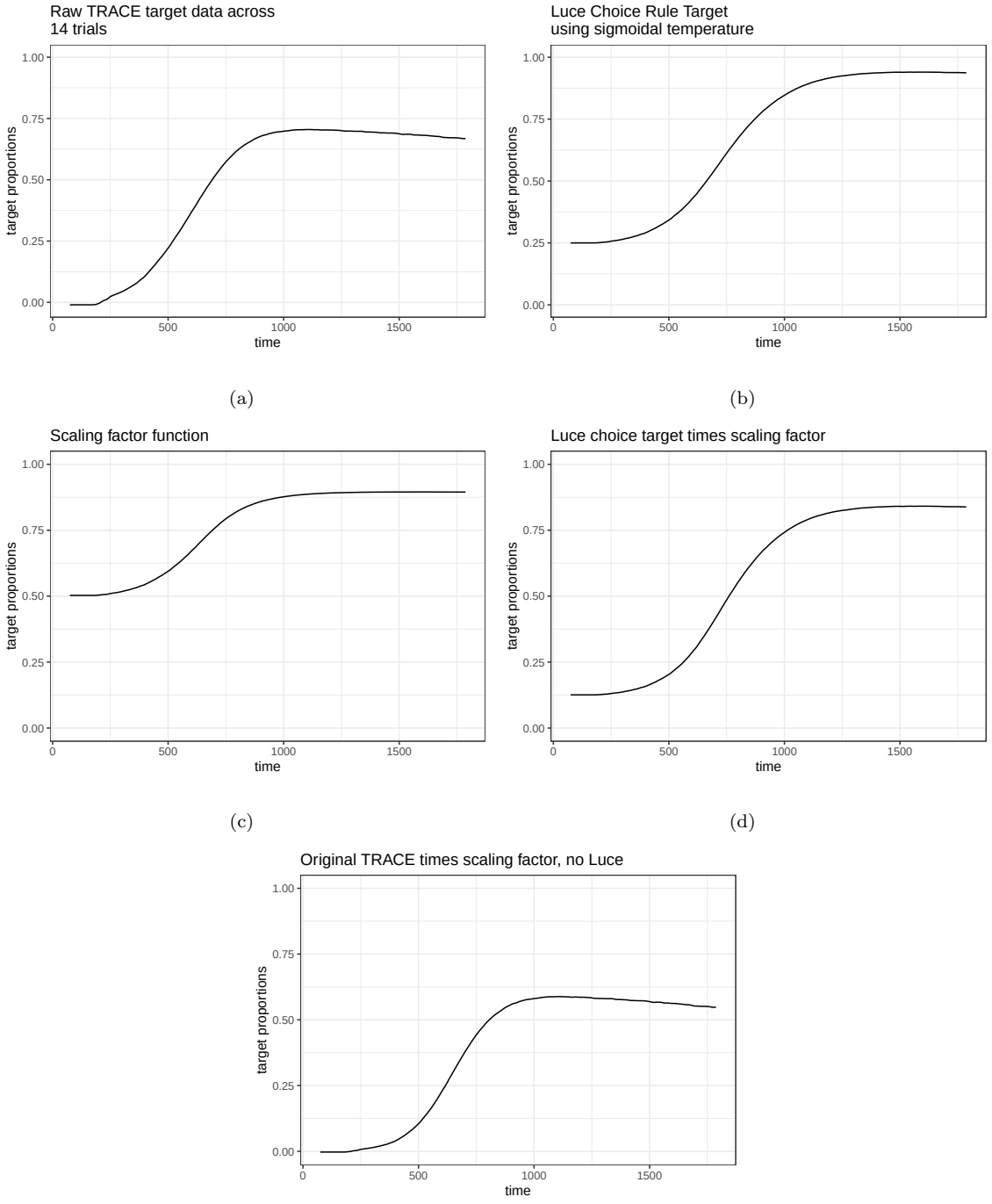


Figure 12: (a) This is simply the raw TRACE data across the 14 simulations with standard parameters. (b) Transformation of TRACE activation using LCR with sigmoidal temperature. (c) Scaling factor function built on max activations *after* performing LCR, using peak/baseline values from target object. (d) TRACE activations following LCR transformation and multiplying by scaling factor. This is what I have been using as model prediction of fixations, though note the baseline value being near 0.15. (e) Perhaps unnecessary, this is simply investigating TRACE activation by the scaling factor but without first conducting LCR. Note that none of these seem to have both the correct baseline and peak values

An interesting aside, though – if we do not make the adjustment to the saccade data where we anchor asymptotic behavior at 0/1, we get a saccade curve bearing less relation to the fixation curve, but with much higher agreement with the set of TRACE curves, in particular with regards to the baseline point and peak. Presumably with some tweaking, it could be made to match even more closely

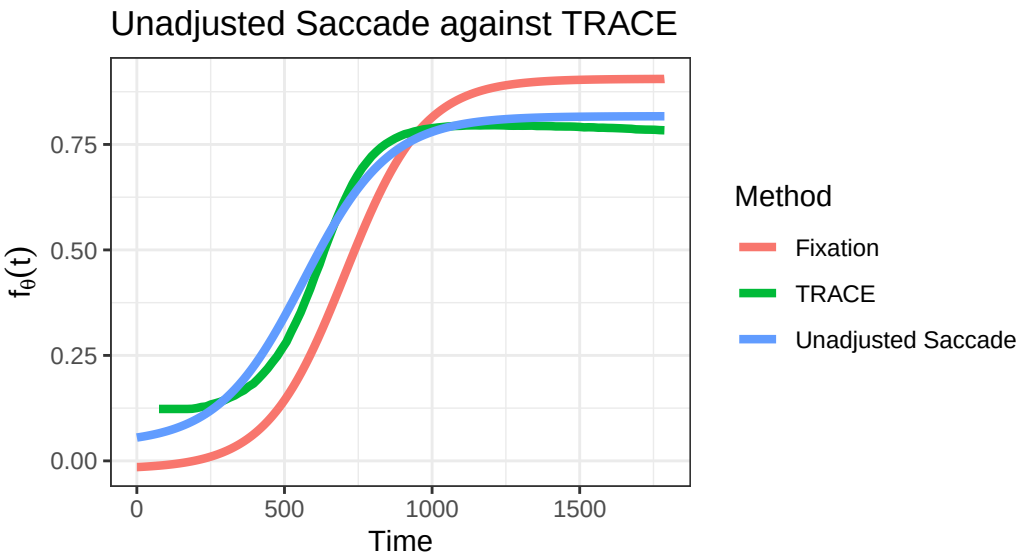
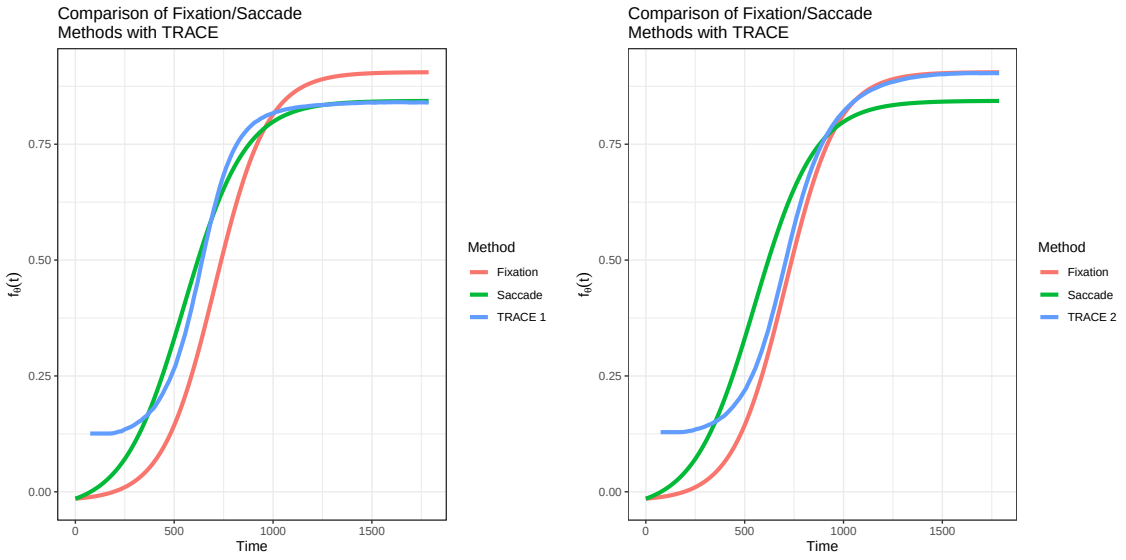


Figure 13: Plot illustarting how the unadjusted saccade method (without anchoring at asymptotes) both matches more closely with the TRACE predictions (particularly near the baseline) while also taking on a far different shape than the fixation curve. This is in contrast to the Princess Bride simulations in which the distortion was minimal and the saccade curve appeared to be more of a horizontal shift

6.3 Comparisons

Here is where I would suggest the consistency of the models. What I show here is a moderated version of this, namely I show that there are two transformations of TRACE (changing the parameters of the sigmoid function for Luce choice rule) that each match one or the other of the fixation/saccade curves better. As such, neither is superior in any sense, but both are in the realm of consistency.



(a) (b)

Figure 14: Examples of different temperatures used in LCR and how this effects TRACE activation. In (a), this leads to greater consistency with the saccade curve; in (b), with the fixation curve. This is evidenced also by RMS error values

Presented in Table 2 is a summary of the RMS error of the 1000s simulations using both the saccade and fixation methods against two instantiations of TRACE. As we see and corresponding to (a) in Figure 14 we have better agreement between the saccade method and trace predictions; this relation is flipped for case (b).

	Method	TRACE	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	Fixation	TRACE1	0.1148	0.1743	0.2181	0.2226	0.2583	0.4407
2	Saccade	TRACE1	0.0749	0.1051	0.1396	0.1449	0.1655	0.2933
3	Fixation	TRACE2	0.0991	0.1270	0.1529	0.1606	0.1712	0.3875
4	Saccade	TRACE2	0.0957	0.1404	0.1830	0.1879	0.2275	0.3734

Table 2: Summary of RMS of two transformations of TRACE against saccade and fixation method

As an aside, this also lends itself to the idea of having a *distribution* of curves associated with lexical activation rather than pursuing point estimation. In some sense, this allows a natural way to account for the observed variability in experimental conditions without having to attempt to model it. Not sure if this is an idea worth elaborating on.

7 Discussion

what have we learned?

Here are really the main takeaways.

8 limitations

probably good idea to keep running list of these all in one place

1. linking hypothesis/cognition curve
2. trace parameters maybe/general degrees of freedom
3. only evidenced on logistic, though for practical not theoretical reasons
4. adding parametric form (necessity for saccade method)
5. oculomotor delay, where to discuss

9 appendices

Here I am just including more or less random sections that either do not have a definite place yet in the main body of the paper, are part of what might be considered future work, or truly are things that belong in the appendix. Presented in no particular order

Appendix A

Treatment of empirical data from McMurray 2010 to get fixation and saccade curves, along with treatment of TRACE data (pending)

Appendix B

I'm not sure if appendix appropriate, but discussion on why double gauss/cohort not considered. This is primarily a consequence of failure to fit adequate models with `bdots`, arising from the fact that `gnls` is highly sensitive to starting parameters. I have demonstrated that they *can* be fit, but successful fits are able to be acquired with a huge range of parameters, bringing into question any validity. As the point of this paper is to demonstrate bias and counter saccade/fixation methods, this seemed an unnecessary addition.

Appendix C

Maybe catch-all for all things OM related. Originally included work showing that fixed delay simply results in horizontal shift, as well as investigation into how the amount of bias is a function both of the length

of delay along with the derivative of the generating curve around the delay. Bias near the asymptotes has minimal impact relative to delay near the crossover point.

I think this would be interesting for future research but a bit beyond the scope to detail much here. Could expand on the idea if there was interest as I already have code written up that samples differentially at different time points.