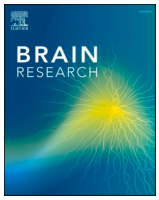
Brain Research 1856 (2025) 149563

Contents lists available at ScienceDirect

Brain Research

journal homepage: www.elsevier.com/locate/brainres



TRACE-ing fixations in the Visual World Paradigm: Extending linking hypotheses and addressing individual differences by simulating trial-level behavior☆,☆☆

James S. Magnuson *

University of Connecticut, Storrs, CT, USA
BCBL. Basque Center on Cognition Brain and Language, Donostia-San Sebastián, Spain
Ikerbasque. Basque Foundation for Science, Bilbao, Spain

A B S T R A C T

I review several alternative linking hypotheses for relating eye tracking data from the VWP to cognitive theories and models. While some models are able to simulate VWP data surprisingly well (such as the TRACE model), there is still ample ambiguity to resolve in the meaning of fixation proportions over time, despite decades of work with the VWP. I also present a simple fixation model based on probabilistic sampling from underlying lexical activation that allows simulation of individual trials. Unsurprisingly, a properly-parameterized sampling procedure approximates the underlying activation patterns when sufficient trials are averaged together. However, the utility of simulating trial-level behavior is not in reconstructing central tendencies (which can be derived directly without simulating fixations), but in addressing, for example, individual differences. I also discuss critiques and misunderstandings of linking models to the VWP, and analogies to a simpler paradigm – lexical decision – to illuminate the logic of linking hypotheses in the VWP.

1. Introduction

The Visual World Paradigm (VWP; Tanenhaus et al., 1995) has been a key tool in the cognitive scientist's kit for 30 years.[1] From many of the earliest VWP papers, an approach that has been particularly impactful on theoretical understanding has been to link the fine-grained time course measure provided by the VWP with fine-grained predictions from computational models.

Spivey-Knowlton (1996) conducted the first simulations of VWP data, modeling the time course of competition between cohorts (words sharing onsets, such as *candy* and *candle*) with his Normalized Recurrence model (see also chapter 7 of Spivey, 2006). Notably, in this early work, Spivey already proposed interaction between visual and auditory stimuli in the VWP (see Spivey, 2025 for an extension of this pioneering work).

Allopenna et al. (1998) took a somewhat different approach, using the TRACE model of speech perception and spoken word recognition (McClelland and Elman, 1986) to model lexical activations, and using lexical activations to feed a decision module based on the visual scene (without feedback from the visual scene to lexical processing). They also looked at an additional type of potential competition, rhymes, such as *candle* and *handle*, which are predicted to compete by the Neighborhood Activation Model (Luce and Pisoni, 1998) but not by the Cohort Model (e.g., Marslen-Wilson and Welsh, 1978; Marslen-Wilson and Tyler, 1980). TRACE unexpectedly provided a close fit to the pattern observed with human participants: early, strong competition (or at least, coactivation; Magnuson, 2019) between cohorts (onset competitors) as well as later, weaker apparent activation of rhyming words. This pattern emerges in TRACE because its pseudo-spectral inputs are presented sequentially over time, in a speech-like manner. This gives the target and

☆ This article is part of a special issue entitled: '30 Years Visual World Paradigm: The State of the Art' published in Brain Research. ☆☆ This research was supported in part by U.S. National Science Foundation grant BCS-PAC 2043903. This research was also supported in part by the Basque Government through the BERC 2022-2025 program and by the Spanish State Research Agency through BCBL Severo Ochoa excellence accreditation CEX2020-001010-S and through project PID2023-149585NB-I00 (ReNeMos). I thank Michael Spivey and Paul Allopenna for extremely helpful comments. I thank Mike Tanenhaus for his mentorship.

* Address: University of Connecticut, Storrs, CT, USA.

[1] Cooper (1974) introduced a similar paradigm, but it went largely unnoticed until after the independent re-invention of the paradigm by Tanenhaus and colleagues. See footNote 1 in Magnuson (2019) for additional discussion. To update some details from that footnote, Cooper had been cited 389 times as of January 13, 2019 according to scopus.com. As of July 6, 2024, it has been cited 577 times (188 more times). However, it is still the case that only 5 citations predate Tanenhaus et al. (1995), and only 6 predate the citation of Cooper by Tanenhaus and Spivey-Knowlton (1996). Thus, dating the contemporary VWP from 1995 – while acknowledging its precedent – is sensible.

https://doi.org/10.1016/j.brainres.2025.149563
Received 23 October 2024; Received in revised form 25 February 2025; Accepted 7 March 2025
Available online 18 March 2025

its cohorts a head-start over rhymes. Activation is governed by lateral inhibition (phoneme-to-phoneme and word-to-word) in TRACE, and this explains why rhymes tend not to activate as strongly as cohorts in TRACE, even when they have higher overall overlap with a target word. By the time the input corresponds sufficiently to a rhyme to activate it strongly, the already-strong activation of the target and its cohorts allows them to inhibit rhymes.

These were watershed developments in the study of human speech processing, as the VWP provided estimates of lexical activations over time that could be compared directly to continuous outputs from computational models. This provided a new window on the dynamic details of speech processing. Subsequent studies used this approach to reconsider unresolved challenges in spoken word recognition. These included using the TRACE model to simulate distinct proposals for the locus of word frequency effects and then testing the predictions with the VWP (Dahan et al., 2001). In another early case (Dahan et al., 2001), VWP data shed new light on a complex experimental paradigm (subcategorical mismatch, where portions of different words and/or nonwords are spliced together to provide misleading coarticulatory cues to upcoming phonemes). Previous studies using a lexical decision version of the subscategorical mismatch paradigm had led to claims that lateral inhibition in TRACE drove predicted response patterns that were in conflict with human data (Marslen-Wilson and Warren, 1994). Dahan et al. (2001) used TRACE activations to simulate both graded effects that they observed in their VWP study (which were highly consistent with the model predictions) and the seemingly divergent lexical decision data patterns (human-like lexical decisions emerged from simulations with an absolute threshold for word responses).

In the three cases mentioned so far (Allopenna et al., 1998; Dahan et al., 2001; Dahan et al., 2001), the VWP was instrumental in providing a basis to adjudicate between theories, and to assess the potential validity of specific computational mechanisms (e.g., lateral inhibition). The timecourse of phonological competition is particularly important. My colleagues and I consider this a hallmark of human spoken word recognition that any candidate theory or model should accommodate (e.g., we reference it as a fundamental prerequisite for model validity in our papers on the TISK and EARSHOT models; Hannagan et al., 2013; Magnuson et al., 2020).

Other teams have compared implemented models to VWP data as well. For example, Norris and McQueen (2008) used their Shortlist B model to simulate aspects of the word frequency studies of Dahan et al. (2001). Another example is Smith et al. (2017), who used their Multimodal Integration Model to simulate their VWP results showing different degrees of apparent competition due to phonological, semantic, or visual similarity. Magnuson et al. (2003) simulated not just the timecourse of phonological competition, but its development as adults learned new sets of artificial words using a Simple Recurrent Network (Elman, 1991; Elman, 1990).

Something all these examples have in common is that they simulated VWP patterns as central tendencies from models, rather than simulating the specific behavior underlying human VWP data: saccades and fixations. From the beginning (Allopenna et al., 1998; Spivey-Knowlton, 1996), central tendencies from models have been conceptualized as (and/or transformed to) quantities like predicted response (fixation) probabilities or proportions. They have relied on linking hypotheses that assume that if actual fixation decisions are generated probabilistically from underlying lexical activations, the processes sampling from lexical activations would generate central tendencies (fixation proportions) strongly resembling those underlying activations. In the next section, I discuss a critique of this approach to modeling VWP data (Norris, 2005) that asserts that modeling central tendencies rather than individual fixations is an exercise in atheoretical modeling without a linking hypothesis. This motivates a discussion of linking hypotheses, and then simple simulations demonstrating how readily a fixation-generation model recovers underlying central tendencies from model predictions. Rather than concluding that there is no need to simulate individual fixations, I discuss potential advantages of doing so (as well as other ongoing efforts to better link VWP behaviors to models).

1.1. *Are conventional simulations of Visual World Paradigm studies fundamentally flawed?*

Despite the apparent successes and progress afforded by linking implemented computational models to the VWP, Norris (2005, , pp. 337-338) argues that the conventional approach introduced by Allopenna et al. (1998) is fundamentally flawed and lacks a proper theoretical basis, and that researchers have mistaken superficially good fits for theoretical insights. I will review this critique in detail as motivation for a deeper examination of linking hypotheses in the VWP – and other paradigms. Several assertions made in the critique are emblematic of common misconstruals in discussing how a linking hypothesis allows a model of internal processing to simulate human behavioral data. By dissecting each misunderstanding in the critique, the goal is to provide greater clarity regarding the components entailed in a linking hypothesis.

As we review this critique, it will be helpful to refer to the eye tracking data and simulation results from Allopenna et al.; Fig. 1 presents the key results. Participants' fixations mapped closely onto phonetic similarity over time with a lag of 200 ms (Fig. 1A). TRACE simulations captured the key patterns (target and cohort separating early from rhyme and unrelated items, with a later peak for rhymes). This emerged in TRACE due primarily to lateral inhibition (Fig. 1B: the bottom-up input initially activates items matching the input from onset (the target and its *cohorts*; when the input favors the target, cohort activations decline due to lateral inhibition from the target; at the same time, the input becomes sufficiently similar to rhymes that they are activated robustly above the unrelated baseline; however, rhymes never reach as high a peak as cohorts because they are already being inhibited not just by the target but also by the cohorts. While there was an obvious qualitative correspondence between TRACE activations and fixation proportions, with a few additional parameters (discussed later), Allopenna et al. converted TRACE activations into predicted fixation proportions Fig. 1C) that provided a very close quantitative fit to the data (as is clear from the small discrepancies plotted in Fig. 1D, and high R^2 and low root mean squared error values calculated by Allopenna et al.).[2]

Now let's consider Norris's critique, starting with a lengthy quote, because the details and full context are critical. Key points to be discussed later are highlighted with bold font.

> Two example *[sic]* of models without a proper theory are the perturbation model of memory (Lee and Estes, 1981; Lee and Estes, 1977) and simulations of eye-movements during spoken word recognition presented by Allopenna, Magnuson and Tanenhaus (1998). The perturbation model simulates serial position curves, mainly in immediate serial recall experiments. In these experiments subjects are presented with a sequence of items (frequently digits or letters) and are required to recall the items in the correct order. The perturbation theory is a statistical model that generates predictions of the probability that each item in a list is recalled at each output position. For example, it simulates the fact that items at either end of a list are more likely to be recalled in their correct position than are items in the middle of a list. The problem is that the model doesn't explain how any particular list is actually recalled. **In effect, the model describes the data, but doesn't describe the behaviour or mechanism that produces the data**.

[2] Note that Allopenna et al. (1998) labeled their observed human fixation proportions as fixation *probabilities*. In later work (including here), we use different terminology: we relate predicted fixation *proportions* to observed human fixation *proportions*.

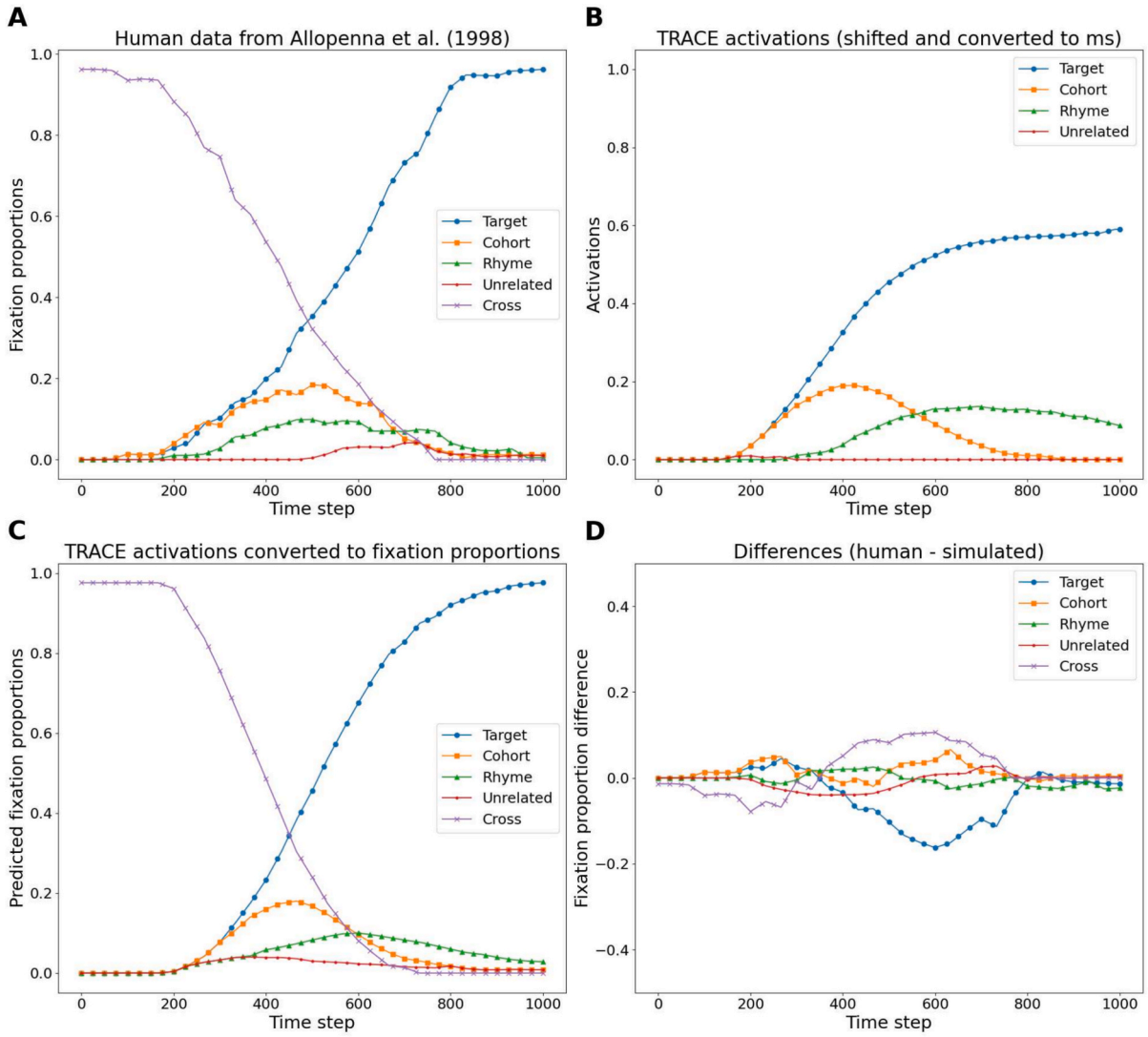


Fig. 1. Key results from Allopenna et al. (1998), Experiment 1. A: Their eye tracking (VWP) data. B: Their reported mean activations by items types in TRACE simulations. C: TRACE activations converted to predicted fixation proportions over time. D: Discrepancies between human data and predicted fixation proportions (A-C). *Panels A-C adapted from figures in the original paper.*.

Allopenna et al. present simulations using TRACE (McClelland and Elman, 1986) **that suffer from exactly the same problem.** They simulated data from an experiment tracking listeners' eye movements to pictures of objects as subjects listened to sentences. The sentences could contain either the names of those objects, or the names of phonologically similar objects. For example, the sentence might be "Pick up the beaker, now put it in front of the diamond" and there might be pictures of a beaker, a beetle, a speaker and a control such as a carriage.[3] Activation of the spoken words was simulated in TRACE. **TRACE can simulate the probability of recognising each word at each moment in time.** Plots of these probabilities show an excellent fit to the eye movement data (i.e. the probability of fixating on the different pictures). **The fit is so good, that surely it must be a good model** (see Roberts and Pashler, 2000; Roberts and Pashler, 2002, for further discussion of the relation between good fits and the quality of models). But, **consider what subjects must be doing whenever the model predicts that they are equally likely to fixate either of two pictures.**

Clearly subjects can only be fixating one of the two pictures at any point, but the model doesn't say which one. In fact, **the model doesn't say how long each picture will be fixated for either.** Predictions at one point in time are completely independent of the following point in time. **The model therefore predicts that, from moment to moment, the subjects' eyes flicker randomly from picture to picture.** Although the model does a very good job at simulating behaviour averaged over trials, **it doesn't give a plausible explanation for what subjects do on an individual trial.** To use terminology that Tanenhaus and Magnuson use elsewhere, **the model doesn't have a linking hypothesis** (Tanenhaus et al., 2000).

These are strong claims. If they are valid, it would appear that the visual world paradigm is not linkable to computational models (or theories) unless the model (and theory?) specifies precisely how individual behaviors (in this case, fixations, in others, perhaps button presses) are generated. Given that Norris and McQueen (2008) employ

[3] To clarify, all trials included an image of the named target item, along with images of three other items. On some trials, those other images included pictures of onset and/or rhyme competitors. If the display included the target and one competitor, the other two pictures were phonologically unrelated to the target. If both the cohort and rhyme appeared, the one other item was unrelated. After the instruction to pick up the target item, the next instruction was to place the target next to, above, or below one of the shapes.

key aspects of the linking hypothesis detailed by Dahan et al. (2001), perhaps no one considers it an important challenge for modeling VWP data. However, such a strong critique merits a direct response, so let us evaluate the arguments. Key points are numbered to facilitate referencing specific arguments.

The first key claim is that **(1)** *The model describes the data, but not the behavior or mechanism that produces the data …simulations using TRACE suffer from exactly the same problem.* This is a surprising assertion. Allopenna et al. clearly set up the theoretical principles at stake in their experiments: they aimed to contrast *mechanisms* posited and/or implied by models like Cohort and the Neighborhood Activation Model with TRACE. The Cohort Model (e.g., Marslen-Wilson, 1987; Marslen-Wilson and Welsh, 1978; Marslen-Wilson and Tyler, 1980) proposes that words are continuously mapped from phoneme sequences to words, and that phonemes in the input provide both positive evidence for words corresponding to the sequence, and negative evidence against words mismatching the input, which should therefore be removed from the recognition 'cohort' (and so only words overlapping at onset are predicted to compete strongly for recognition), with a relatively hard 'reset' when word boundaries are detected (e.g., when an incoming phoneme cannot be added to the sequence buffer to continue adding a word, as in/kætd/ which could be part of the phrase *the cat drinks*; in this case, *cat* would be recognized and/kæt/ would be removed from the phoneme sequence buffer, leaving only/d/ – the presumed onset of the next word – until more input arrives). Thus, only words overlapping at onset are predicted to compete strongly for recognition. The Neighborhood Activation Model (NAM; Luce and Pisoni, 1998) provides a very different proposal. NAM does not consider segmentation (it only considers isolated or segmented words) nor *where* two words differ from each other (with the effect that two words are predicted to compete strongly so long as they mismatch by only one phoneme, no matter the position of the mismatch). Thus, NAM assumes word boundaries are known, as only entire wordforms are considered.

In contrast, *continuous mapping models*, like TRACE, do not explicitly track word boundaries, and instead allow words to become activated at any point based on match to bottom-up input. TRACE turns out to make predictions intermediate between Cohort and NAM: onset (cohort) competition is early and strong (e.g., *candle*, *candy*), while neighbors mismatching at onset still compete, but with a later and lower peak (e.g., *candle*, *handle*). Thus, competitors are not limited to words overlapping at onset, but those items compete more strongly due to a temporal advantage: the onset competitors (including the target) are activated early by bottom-up match. Rhymes match later-arriving inputs, but cannot be as strongly activated because they are inhibited by the onset competitors (especially the target; see Allopenna et al., 1998, pp. 425-426). Thus, it is not the case that the predictions from the TRACE model are not mechanistically linked to the behavioral data; the specific patterns observed follow (non-intuitively) from interactions of feedforward, feedback, and lateral (inhibitory) connections (as mentioned earlier).

(2) *TRACE can simulate the probability of recognising each word at each moment in time.* While this would arguably be possible to simulate, *'probability of recognition'* is not a reasonable interpretation of TRACE activations. One could take raw activations or response probabilities derived from the Luce choice rule and simulate something like a forced response to a 'gated' (partial) input (as Allopenna et al. did to simulate gating in their second experiment). In that case, the linking hypothesis to the task would indeed operationalize response probability as the probability of *responding* with one word or another (e.g., as human subjects are asked to do in the gating task, giving completions such as 'beam', 'beaker', or 'beetle' in response to the input /bi/). But even in this case, it would not be the probability of *recognizing* each word at that moment in time; it would be proportional to the relative probability of each word given the available evidence. This may seem like a minor misunderstanding, but it is fundamental. Interactive activation models like TRACE eschew the idea of a 'magic' moment of recognition (Balota, 1990). The models generate activations over time, and it is up to the researcher to operationalize recognition based on activations.

(3) *The fit is so good, that surely it must be a good model.* Here, Norris implies that Allopenna et al. made such a claim, or that they presented the model fits without examining their theoretical import. However, Allopenna et al. make no such claims or assumptions. Roberts and Pashler (2000) point out that while theories with corresponding models 'gain credence' when they provide good fits to data, good fits alone are not deeply meaningful. One must also consider the flexibility of a theory and corresponding model (what the model *cannot* fit), what patterns the model is known to generate (e.g., whether the model generates patterns not observed in human subjects, or fails to generate patterns observed in human subjects), or whether the model can simulate *any* plausible result (e.g., patterns predicted by incompatible theories). Roberts and Pashler (2002) put it like this: 'that a theory fits data is meaningful only if it was plausible that the theory would not fit.' Let us consider these points. TRACE does not predict just *any* outcome whatsoever: it does not predict equivalent activation of any close (1-phoneme mismatch) competitors, like NAM does; it does not predict that only words overlapping at onset compete, like Cohort does. On the other hand, it predicts rhyme effects should be absent in the gating paradigm (Allopenna et al., 1998, Experiment 2), providing an unexpected explanation for why Marslen-Wilson and colleagues observed so many results consistent with the Cohort model and apparently incompatible with NAM using that paradigm.

Other data also supported Cohort over NAM. For example, Marslen-Wilson and colleagues have found cross-modal semantic priming consistent with Cohort (*beaker* would prime *insect*, suggesting hearing *beaker* sufficiently activates *beetle* for semantic priming of its relative *insect*. However, *beaker* would not reliably prime *stereo*, a semantic associate of its rhyme, *speaker* (Marslen-Wilson and Zwitserlood, 1989). Since TRACE does not have semantic representations, it is not possible to simulate this result directly. But as Allopenna et al. noted, it is plausible that if underlying activations are as predicted by TRACE, rhyme activations could be sufficiently weaker than cohort (onset) activations that they could not drive detectable spreading activation to semantic representations (although this possibility remains untested). Finally, was it plausible that the model could have *failed* to simulate these results? Indeed; it was not intuitive (except perhaps to Paul Allopenna) that the observed pattern of results should emerge in TRACE prior to running the simulations. Notably, McClelland and Elman (1986) emphasize that their intent was to implement a model consistent with the principles of the Cohort model. While Allopenna et al. are careful to avoid drawing strong conclusions for the TRACE model specifically (p. 431), their results clearly support continuous mapping models more generally (although TRACE continued to simulate well even more subtle manipulations, e.g.: Dahan et al., 2001; Dahan et al., 2001).

Comparing Fig. 1A and 1B qualitatively, we already observe tremendous progress in interpreting the eye movement pattern; the unanticipated competition patterns emerge naturally from TRACE. Allopenna et al. went a step further, asking whether a few assumptions about eye movements could yield a *linking hypothesis* that might explain finer details of the data. The fundamental idea was that lexically-driven eye movements depended on at least two component processes: lexical activation, and then a decision process that guided eye movements. Allopenna et al. modeled this by allowing all items in the lexicon to compete, but then estimating the outcomes of a separate decision process that would only consider the four possible choices (the four pictured items) without sending any feedback to the lexical processor (TRACE).

They made three key assumptions. First, they needed to map cycles in TRACE to real time (ms) in the eye movement data. They did this by simply calculating the mean ms per phoneme in their items and cycles per phoneme in TRACE, and transformed cycles to ms with a simple constant (with each TRACE cycle equal to approximately 11 ms). Second, given an approximate lag of 200 ms between phonetic details and changes in fixation proportions (approximated based on a lag of

approximately 150 ms for saccades in extremely simple tasks; e.g., Matin et al., 1993), they simply shifted TRACE activations by 200 ms (as in Fig. 1B). Third, at each time step, they calculated predicted fixation proportions using the Luce Choice Rule (LCR; Luce, 1959). We will examine the LCR in more detail below; for now, note that it applies a nonlinear normalization that amplifies larger values and squashes smaller ones, and was used to model choice behavior in TRACE in prior work (McClelland and Elman, 1986).

Allopenna et al. also applied two additional parameters. First, there is a constant parameter in the LCR, *k*, that determines the degree to which values will be amplified or squashed. They made *k* change over time via a sigmoid function (such that *k* was minimal in the early time course, and rapidly transitioned to a maximum value midway through the timecourse). This improves fit by damping later predicted rhyme fixation proportions. Second, they scaled response probabilities based on the ratio of the maximum underlying activation value at a time step to the maximum activation value observed at any time step. This allowed initial response probabilities to be (near) zero, without explicitly modeling fixations to the central fixation cross (cross fixation proportions were added to Fig. 1C as $1 - \Sigma p$, the sum of predicted response proportions to the four objects, to enhance comparability to the eye tracking data).

Thus, the fact that the fit was so good is meaningful. Again, understanding how the underlying activation pattern of phonological competition over time emerges in TRACE is the core theoretical advance. The improved fit that resulted from the linking hypothesis was also logically motivated, and was not a simple exercise in data fitting. That said, four of the five parameters described above *were* free parameters that were adjusted to improve fit (note that the scaling factor for mapping cycles to ms as this was not a free parameter, but a value dictated by the stimulus materials):

1. *k*: scaling factor in the LCR
2. *x*: steepness parameter for Allopenna et al.'s 'dynamic' *k*
3. Δt: scaling factor (ratio of maximum *activation* value at a time step relative to the maximum activation at any time step
4. *Temporal shift*: how far TRACE activations were shifted later in time[4]

These parameters were not exhaustively explored, but were settled upon with trial and error. Nonetheless, the quantifiably good fits reported by Allopenna et al. were not just an exercise in data fitting. Indeed, later papers (Dahan et al., 2001; Dahan et al., 2001) dropped the dynamic *k* adjustment (thus dropping *x* and Δt, while also using the same value for *k* [7], essentially removing it as a free parameter, too), because they modified the paradigm such that participants were equally likely to be fixating any item at time 0. As we shall see below, we can also drop them with Allopenna et al.'s data with a slightly modified linking hypothesis, deflating completely any claim that the linking hypothesis is just data fitting.

(4) *The model predicts subjects' eyes flicker randomly between pictures.* This assertion is based on the following points: (4a) *"The model predicts equal probability of fixating two pictures at a given time point, but does not specify which one is fixated."* (4b) *"…The model doesn't say how long each picture will be fixated for either. Predictions at one point in time are completely independent of the following point in time."* Thus, according to Norris, (4c) *"The model therefore predicts that, from moment to moment, the subjects' eyes flicker randomly from picture to picture."*

However, 'predicted fixation probabilities' does not have these implications. This may stem from a confusion between saccade probabilities (probability of an eye movement at a particular time step) and fixation probabilities (probability of gaze to an item at a particular time step) arising from Allopenna et al. using terms like 'fixation probabilities' rather than 'fixation proportions'. However, response probabilities are a common currency in the psychological, cognitive, and neural sciences. Consider a simple two-alternative forced choice (2AFC) task, such as lexical decision. On each trial, the participant hears a token, which may be a word or a nonword (e.g., *stack* vs. **clush*). On word trials, participants sometimes answer 'no' erroneously (perhaps due to uncertainty, lack of lexical knowledge, motor errors, or noise in the response system, etc.). Over a set of many word trials, we can estimate the probability of participants responding 'yes' or 'no' over time in response to word stimuli, as in the hypothetical estimates in panel A of Fig. 2. Here, the green 'yes' line indicates that participants make few responses in the early time course, many more in the middle, and few in the late time course (because they have usually already responded). The dashed black line indicates that there is also some likelihood that participants will make incorrect rejections and respond 'no' in our hypothetical word trials. Now consider the early time range where the red and green lines overlap (e.g., around the time point indicated by the black circle). Does this indicate that at that point, participants were simultaneously responding 'yes' and 'no', or that their fingers were somehow instantaneously flitting between response options? No; this simply indicates that 'yes' and 'no' responses were made on *equal proportions of trials* during that time range. That is, we estimate the probabilities (calculate proportions) by aggregating over trials (while of course each lexical decision trial includes just one response, whereas multiple fixations may occur in one VWP trial).

The same principles hold for VWP data (Fig. 2, panel B). We most commonly structure VWP data as proportion of fixations over time to categories of items. These *observed* response proportions are our best estimate of the likelihood that participants will, on average, **be** *fixating* a particular item over time – not the probability of a saccade. Thus, just as observed response proportions do not indicate that participants' eyes were 'randomly flickering' between items (they represent simply the proportion of trials where subjects were fixating each item at each time point), predicted fixation proportions are not predicted saccade launches.

That said, how we link them to computational models – how we derive, for example, *predicted* fixation proportions over time – is an important consideration. In the next section, I walk through the logic of data aggregation in lexical decision and visual world paradigms, and then the logic of linking model predictions to those data at various grains (from individual trials to condition or category means). I will return to flaws in points 4b and 4c there, after discussing the final key point of the critique.

(5) *The model lacks a linking hypothesis.* Norris asserts that because Allopenna et al. did not simulate individual fixations, "to use terminology that Tanenhaus and Magnuson use elsewhere, the model doesn't have a linking hypothesis." At this point, Norris cites Tanenhaus et al. (2000), which was a response to a theoretical paper about feedback in models of spoken word recognition, but almost certainly intended to cite Tanenhaus et al. (2000). Much of the latter paper is devoted to discussions of the logic of visual world linking hypotheses, including detailed discussion of the specific linking hypothesis developed by Allopenna et al. (pp. 567–570) and applied (with simplifications) subsequently by Dahan et al. (2001) (pp. 570–573) and Dahan et al. (2001) (pp. 573–576). To assert that the approach lacks a linking hypothesis without acknowledging or discussing the detailed consideration Tanenhaus and colleagues devoted to this topic is surprising.

It seems that the claim here is that linking hypotheses must be at the grain of the exact behavior generated by human subjects. Thus, if we wish to compare a model to lexical decision data, we must simulate individual trials with operational definitions that can generate 'yes' and 'no' decisions for words and/or nonwords. This raises an interesting question: must computational models simulate *individual trials* to be

[4] Some readers may consider this temporal shift a fixed parameter, since virtually all simulation works uses a shift of 200 ms. However, this *was* a free parameter in Allopenna et al.; the standard might have been a bit smaller or larger had that improved the fit substantially. In later work that has adopted the 200-ms value, this has become a fixed parameter.

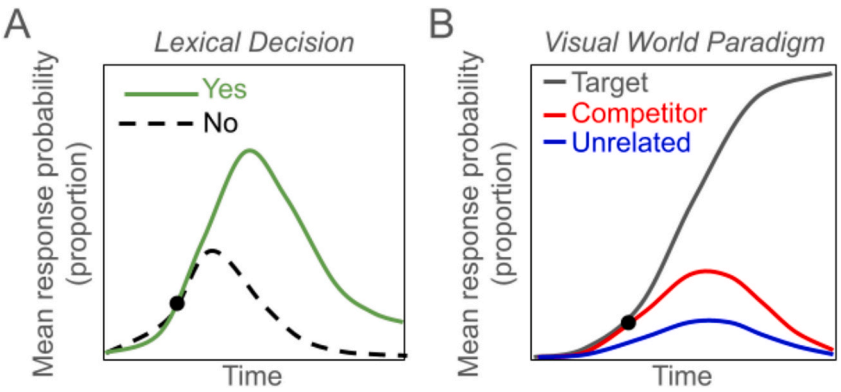


Fig. 2. A: Hypothetical proportions of responding 'yes' and 'no' over time on word trials in a lexical decision experiment. B: Hypothetical proportions of fixating different categories of items over time in a visual world eye tracking study. Both panels show central tendencies and provide an empirical basis for estimating response probabilities over time. Black circles highlight time points where equivalent probabilities are observed for two different responses. *Reproduced from* Magnuson (2024)..

linkable to human data, or can we productively link models and data at the level of central tendencies, such as condition means? Consider two examples where the latter approach has been applied. First, Norris et al. (2000) use their Merge model to simulate lexical decision and phonemic decision data (see their Figs. 2 and 3) from a subcategorical mismatch paradigm (Marslen-Wilson and Warren, 1994; McQueen et al., 1999). They compare condition means from the model (time for the target node to reach a threshold level of activation) – not simulations of responses (e.g., yes/no decisions) to each stimulus – to mean human response times in those conditions. On the logic of Norris (2005), would this be a case of a "[model] without a proper theory", since the simulations use condition means rather than simulations of individual trials? Of course, they had to simulate individual trial *activations* to obtain the condition means – as did Allopenna et al. to obtain their predicted mean fixation proportions over time. But they do not report whether they actually simulated individual *responses* or compared condition means. I discuss this subtle distinction below, in the context of Fig. 3.

Second, Norris and McQueen (2008) present simulations of the VWP data from Dahan et al. (2001) using their Shortlist B model. They compare central tendencies (predicted response probabilities over time) to mean fixation proportions over time – just as was done by Allopenna et al. (1998, 2001,), and others. Although they cite Norris (2005), albeit for a different topic, they do not address the fundamental flaws Norris (2005) claims this approach has. Indeed, they explicitly cite and adopt aspects of the linking hypothesis of Dahan et al. (2001) (e.g., relating model time to real time by calculating msecs per phoneme in real stimuli; shifting response probabilities later in time to account for the fact that fixations lag changes in phonetic details by 200 ms; and calculating final probabilities relative to the types of items in the display, rather than relative to the entire lexicon).

It also bears mentioning that many useful comparisons of models and humans do not simulate individual trials. Even in studies of word recognition that examine item-specific data (such as accuracy or RT), processing of a particular word will typically be simulated just once, and then compared to the mean human performance for that word based on averaging several individuals' trials for that item. In other cases, item-specific predictions are not of interest, but the impact of specific factors, such as word frequency or neighborhood size, are the focus. In these cases, a fairly weak linkage between simulations and data may suffice. For example, it is not uncommon to assume that relatively lower activation for one condition vs. another is compatible with lower accuracy and/or slower response times. However, as a general principle, models that provide greater *specificity* should be preferred: it is commonly advised that modelers should prefer models that provide quantitative fits (vs. qualitative fits), and prefer models that provide item-specific error or timing predictions (vs. central tendencies, etc.; cf. Jacobs and Grainger, 1994; Magnuson et al., 2012).

Taking the preference for greater specificity to an extreme, we could try to simulate details of individual VWP trials, to the level of individual saccades and fixations. For example, we could attempt to accurately simulate fixation distributions for specific words. However, what theoretical utility would this have? It is not immediately apparent that any theoretical questions hinge on this grain of analysis (though it is possible that probing this grain could reveal important constraints on or differences between models, such as the impact fixations themselves could have on lexical activations; see Spivey, 2025). McMurray (2023) presents a tour-de-force approach to simulating individual fixations, but with a focus on evaluating the validity and reliability of recovering the functional forms of processes underlying eye movements (a goal shared here, as I discuss below). I do not believe we have empirical targets from human data for such a comparison. While one could characterize distributions of fixation parameters for individual words (e.g., probabilities of fixations at different points in time, mean number of fixations to and away from the target image per trial, mean fixation durations, or curve parameters as in McMurray, 2023, etc.), to my knowledge, no one has characterized word-specific parameters with respect to individual differences.[5]

A useful alternative might be to characterize the central tendencies for different words, i.e., the mean fixation curve for a particular item, perhaps when presented among unrelated distractors. One could use Growth Curve Analysis (GCA) extended to the specific time scale and other aspects of the VWP (e.g., Magnuson et al., 2007; Mirman et al., 2008; Mirman, 2014), and use, e.g., intercept, slope, and quadratic terms to characterize the central tendency for that word (for examples of individual differences approaches using GCA, see: Mirman et al., 2008; Mirman, 2014). One could generate simulations and similarly characterize predicted curves (either based on fixation simulations as I explore below, or by taking activations or response probabilities directly). Then models could be compared on their ability to capture word-specific differences in GCA parameters (as in Mirman et al., 2011), or other parameters characterizing curves, as on the BDOTS approach (Oleson et al., 2017). Whether stable, systematic differences would emerge for VWP responses to individual words is an open question. Addressing this issue would be exploratory at first, but could be quite productive. For example, one could examine sets of words that are highly similar in

[5] There are interesting dynamics that arise from factorial groupings (e.g., differences in the impact of high- vs. low-frequency competitors, Dahan et al., 2001, , which the authors simulated with TRACE), and even more subtle differences that arise due to fine-grained phonetic details as a function of individuals' phonological skill (Li et al., 2019, , though the authors did not conduct simulations). These examples get closer to the level of individual items.

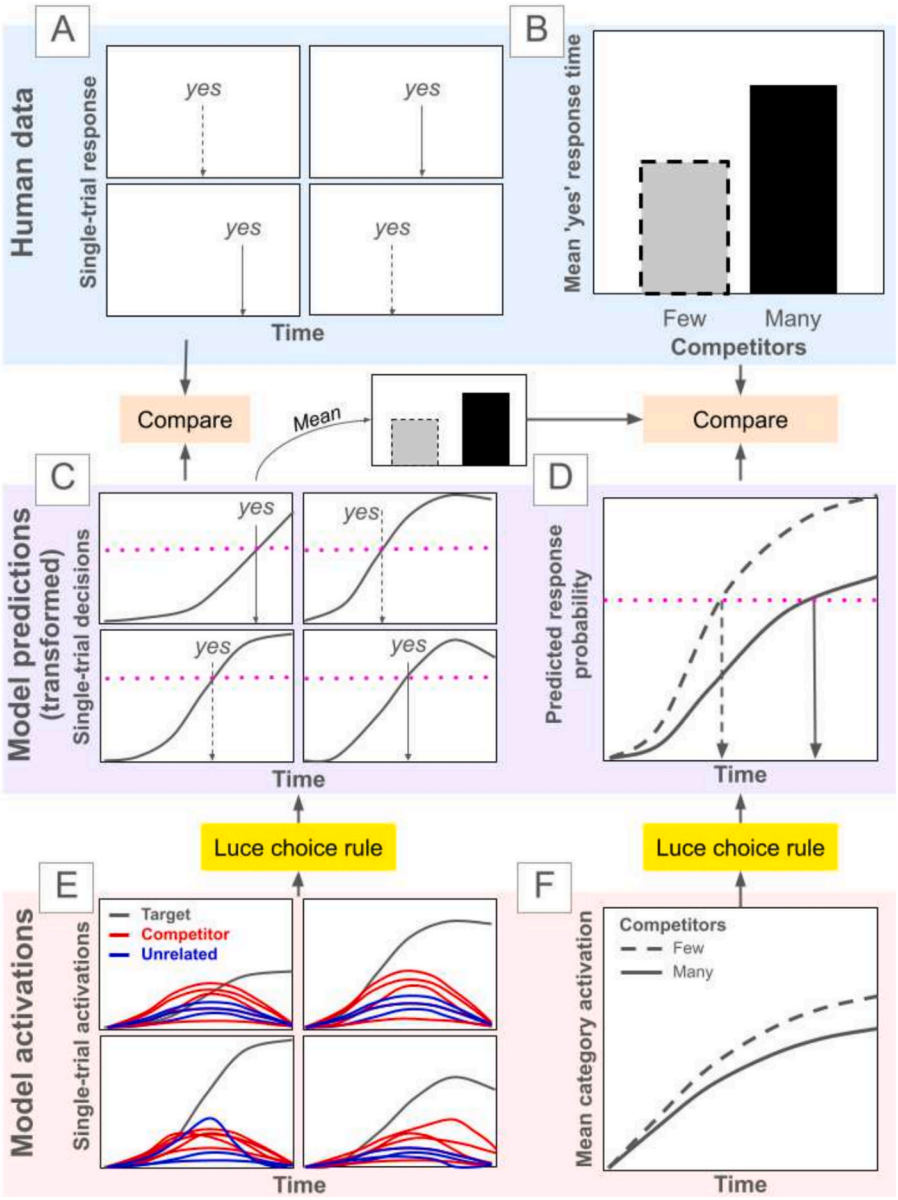


Fig. 3. Simulating lexical decision. From individual trial data (human behavior) (A), we typically focus on RT means by condition (B, derived from correct 'yes' responses). Models typically begin with simulations of individual words (E). We operationalize 'yes' ('word') responses by transforming activations (E) to response probabilities (C); a 'word' response occurs if/when the target probability reaches a threshold, allowing us to compare model and human means ('compare' path from C to B). Or we could assess word-specific predicted RTs ('compare' path from C to A). Alternatively, we could calculate mean target activations for items with Few or Many competitors (F), convert them to response probabilities (D), and apply a threshold. *Reproduced from* Magnuson (2024)..

terms of key parameters such as frequency, length, and number of competitors. If there were substantial variation in the GCA parameters among those words, one could then attempt to discover quantifiable differences between them – in phonotactic probability, concreteness, semantic density, etc. This could inform theories regarding characteristics of words that influence recognition facility.

This discussion raises interesting questions. Can a fixation mechanism stochastically sampling from model activations easily recover central tendencies (as Allopenna et al., 1998, assumed), or are there unforeseen challenges? Even if adding a fixation module results in similar central tendencies in predicted fixation proportions, might simulating fixation behaviors provide unexpected insights? In the next section, I lay out linkages between models and individual responses (button presses or fixations) vs. central tendencies for lexical decision and VWP tasks. My aim is to make these linkages as clear as possible before moving on to simulations of individual fixations.

2. Linking models and tasks: Two examples

Let's begin by considering how we can link model simulations to a lexical decision task. Lexical decision is typically a two-alternative forced choice (2AFC) task, with one button for 'yes' (the item is a word) and another for 'no' (the item is not a word). In Fig. 3, the top row schematizes human data (raw and aggregated), the bottom row schematizes lower-level model outputs, and the middle row bridges between the two. The top two panels (A and B) give a schematic of individual trials and central tendencies. If we are interested in the impact of number of competitors on correct 'yes' response times, we would aggregate the single data points from individual trials (A) to calculate condition means (B). The challenge is relating simulated lexical activations (E) to the human data. One possibility is to average together target word activations for the two conditions (F). We could then set a threshold, and determine when, on average, words exceed that

threshold, and those points become the predicted mean RTs for our two conditions.[6] We could do this directly on activations, though more commonly, with a model like TRACE, we would first apply the Luce Choice Rule (LCR; Luce, 1959), then apply the threshold (following the path from panel F to panel D), and then compare the model means to the human means (path from panel D to panel B).

Alternatively, we could simulate individual trials. In Fig. 3C, this is schematized as applying the LCR to individual trial activations (from panel E). We then apply a threshold to each target activation to get item-specific RTs, These could be directly compared to item-specific RTs from human data (A), or they could be averaged and compared to condition means (B). If we worked with raw activations and did not apply the LCR, we could be confident that the condition means we would get from applying a threshold to individual trials (C) or mean activations in different conditions (D) would yield similar means (assuming we use the same threshold).

Applying the LCR entails some additional complexities and complications. First, if we took mean target activations for conditions from the model (panel F), we could not apply the LCR directly; as the label implies, there must be a *choice* to be made. To apply the LCR, we could, e.g., calculate the mean maximum activation of non-targets at every time point. The LCR would then transform differences between targets and non-targets (making predicted response probabilities 0.5 when values are identical, and amplifying differences as the target diverges from non-targets). Also, I stipulated earlier that we are considering correct 'yes' responses. But of course, humans make errors, occasionally responding 'no' to words or 'yes' to nonwords. We can easily restrict the data just to cases where items were words, but then we must exclude trials where humans responded 'No.' We need to do something analogous for our simulations. To do so, we would actually need to apply a response threshold to individual trials. Then, a clear 'yes' trial would be one where the target and no other item exceeds the threshold, and a 'no' trial would be one where no item exceeds the threshold. A third case would be a trial where an item other than the target exceeds the threshold (whether the target does or not). Note that if this occurs for a human participant, we would not know it. For example, if someone pressed 'yes' because the first syllable of CHERRY activated CHAIR very strongly, we would count this as a correct 'yes' for CHERRY. Thus, it would be reasonable to do the same for the model: so long as any word crosses the threshold on a trial the model would make a 'yes' response, with the time of the first word crossing the threshold the RT for that trial.

Now let's consider the linkage for VWP data, consulting Fig. 4. Again, the human data is schematized in the top row, lower-level model outputs are in the bottom row, and the middle row bridges the two. Now, single trials of human data (A) have one or more fixations. On some trials, we observe fixations to competitors and/or unrelated items, and on nearly all trials there will be at least one fixation to the target (but not always). We aggregate the trial-level data, typically averaging over items and then over participants, to arrive at mean fixation proportions (B). The lowest level model outputs are schematized in panel F, where we see activations for targets, competitors, and unrelated items. We can average over items to get mean activations (G) for each category of activation we are interested in (in this example, targets, competitors, and unrelated items). Again, this is the typical approach, and then commonly mean activations are transformed to predicted response proportions over time.

Alternatively, we could add a fixation generation model. Note that this *could* have no impact whatsoever on the underlying model (just as the LCR component had no impact on the underlying word recognition model in many previous papers, e.g.: Allopenna et al., 1998; Dahan et al., 2001; Dahan et al., 2001; Hannagan et al., 2013), if there is no feedback between the fixation module (or a model of the visual aspect of the task) and the word recognition model. However, see Spivey (2025) for a compelling case that there is such interaction, as well as his model that can simulate both eye-tracking and mouse-tracking VWP data. For current purposes, we will assume a fixation module that is appended to the word recognition model – one that does not send feedback to the word model (as in Fig. 4 panel E). Now we would actually simulate fixations on individual trials. Even if the activation patterns for a specific word are deterministic and therefore constant (as in a variant of TRACE where there is no internal or external noise), if fixations are generated probabilistically from activations, we may still observe substantial variability in the exact timing and targeting of individual fixations.

Once we have simulated trial-level fixations, we can aggregate them exactly as we would aggregate human fixation data to arrive at predicted fixation proportions over time (panel C). These could be compared to human data (panel B) as well as to fixation proportions calculated directly from activations (typically with the addition of the LCR, as in panel D) to assess what we might gain by simulating individual fixations.

Now the fixation model will have to have its own parameters. Minimally, we would need parameters that would (a) govern when a saccade might be triggered, and, (b) once a saccade is made, how much time should elapse before another saccade becomes likely, as well as (c) selecting fixation locations. Going back to the claims of Norris (2005) (that models cannot be linked to models without simulating individual responses, such as saccades and fixations), the crucial question is whether applying a fixation generation model (panels F to E) would result in qualitatively different results (C) vs. transforming activations directly into predicted fixation proportions over time (D). This is the focus of simulations presented in the next section.

3. Fixation generation model

There are at least two components to generating fixations: deciding *whether* to fixate and then *where* to fixate. McMurray (2023) takes a complex approach to simulating individual fixations in the VWP. On his approach, subjects are assumed to differ in parameters governing fixation generation (e.g., in one of his simulations, a 4-function sigmoid equation with terms for intercept, maximum, slope, and crossover), and target and competitor fixations are simulated separately (a convenient simplifying assumption). Here, I take a radically simpler approach. First, *whether* to fixate is decided as a function of two parameters: h (*hysteresis*) and s, an exponent that governs how the probability of a saccade accelerates over the hysteresis window. Conceptually, the goal is to have a process that cannot generate a new saccade immediately after another. The h parameter defines a time window over which a new fixation can be made (i.e., the maximum 'wait' time). We want the probability of a saccade to become greater the longer the time from the previous saccade (which is where s comes in). For this to make sense, let's look at the basic equation for determining saccade likelihood given time from the last saccade.

First, at each time step t, we calculate the time since the last fixation onset, Δt (simply the number of time steps that have occurred since the last saccade). Then, we calculate the probability for a new fixation at time t by:

$$P_{\text{fixation}}(t) = \left(\frac{\Delta t}{h}\right)^{s} \quad (1)$$

So when $s = 1$, the probability grows additively. If $s > 1$, we impose a convex bowing to the function (slower in the early portion of the hysteresis window, accelerating later). If $s < 1$, we impose a concave bowing (faster in the early portion, decelerating later). These relationships are shown in Fig. 5.

Now during the simulation of a specific word, to determine at time t whether or not to generate a fixation, we generate a random number between 0 and 1; if that number is less than the value we get from Eq. 1,

[6] Simple threshold models may not suffice, of course, as behavior in lexical decision tasks may be driven by other internal variables, such as familiarity or total lexical activation (e.g., Grainger and Jacobs, 1996).

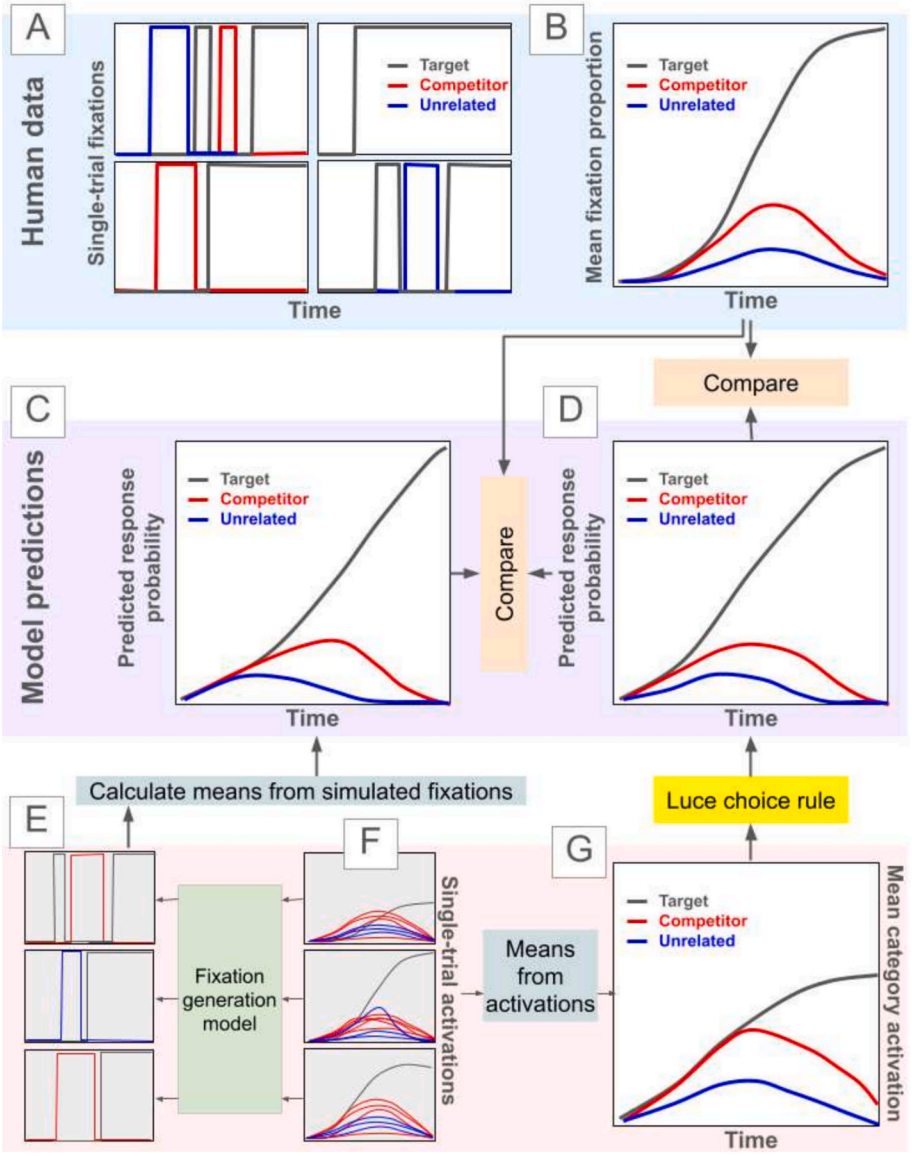


Fig. 4. Simulating VWP data. Single-trial VWP data resembles A, as a human fixates one item at a time. VWP data is aggregated as mean fixation proportions (B). We simulate individual trials (F) and average these (G) and then apply the Luce choice rule to obtain predicted response (fixation) proportions over time (D), which can be directly compared to human VWP means (D to B). Alternatively, we could implement a fixation-generation model that could probabilistically simulate fixations in single trials (E). We could aggregate simulated single-trial data to condition means (C) just as we do with human fixation data (A to B). We could compare these means to human means (C to B) or to the activation-based means (C to D). *Reproduced from* Magnuson (2024)..

we will make a new saccade, and if not, we maintain the current fixation target until the next time step. However, there is a complication here. If we repeatedly make this choice as time progresses from the previous time step, a positive choice will be very like to happen early in the hysteresis window. In fact, with h set to 400 and s set to 2, the probability of a new fixation reaches 0.95 around 110 time steps, and 0.99 around 130 – very early in the acceleration portion of the curve. To avoid this, we need to rescale the curves in Fig. 5 such that the *cumulative* probability of a fixation is exactly 1.0 at time step h. To do this, we calculate a rescaling factor, θ, as follows (where we iterate over all time steps, t, from 1 to h):

$$\theta = \left(\sum_{t=1}^{h} \left(\frac{t}{h}\right)^{s} \right)^{-1} \tag{2}$$

Now we rescale Eq. 1 as:

$$P_{\text{fixation}}(t) = \theta\left(\frac{\Delta t}{h}\right)^{s} \tag{3}$$

This ensures that the cumulative probability sums to 1. The rescaled probabilities are shown in Fig. 6. If we then calculate the cumulative probabilities of fixations for each combination of h and s, we obtain exactly the values in Fig. 5.

So again, if the random number generated at time t is less than the value calculated from Eq. 3, *a fixation decision is triggered*. A fixation location is based on values provided for each possible object in the set *Target, Cohort, Rhyme, Unrelated,* and *Cross*. These *could* be raw activation values, but typically, we rescale activations to response probabilities using the Luce Choice Rule (LCR). First, we have to derive values for the Cross (we only get activations for the four pictured items from TRACE). A simple way to do this is to define σ_c as the summed activations of *Target* + *Cohort* + *Rhyme* + *Unrelated* at cycle (TRACE time step) c. Then we can find σ_{max}, the maximum summed activation value

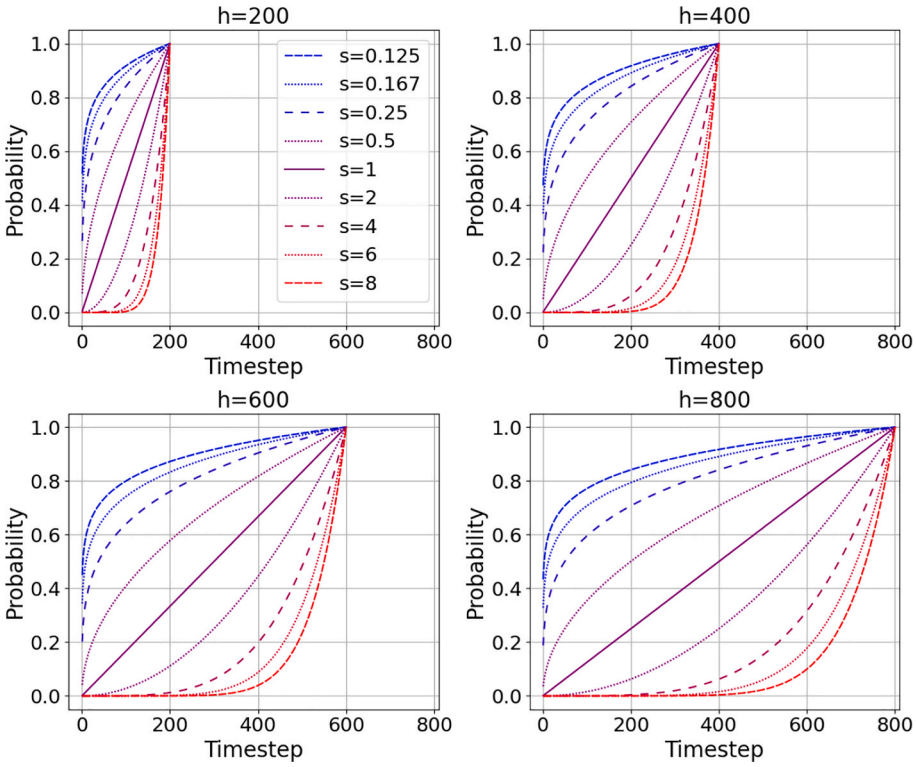


Fig. 5. Relationship between hysteresis (h) and the s exponent in Eq. 1.

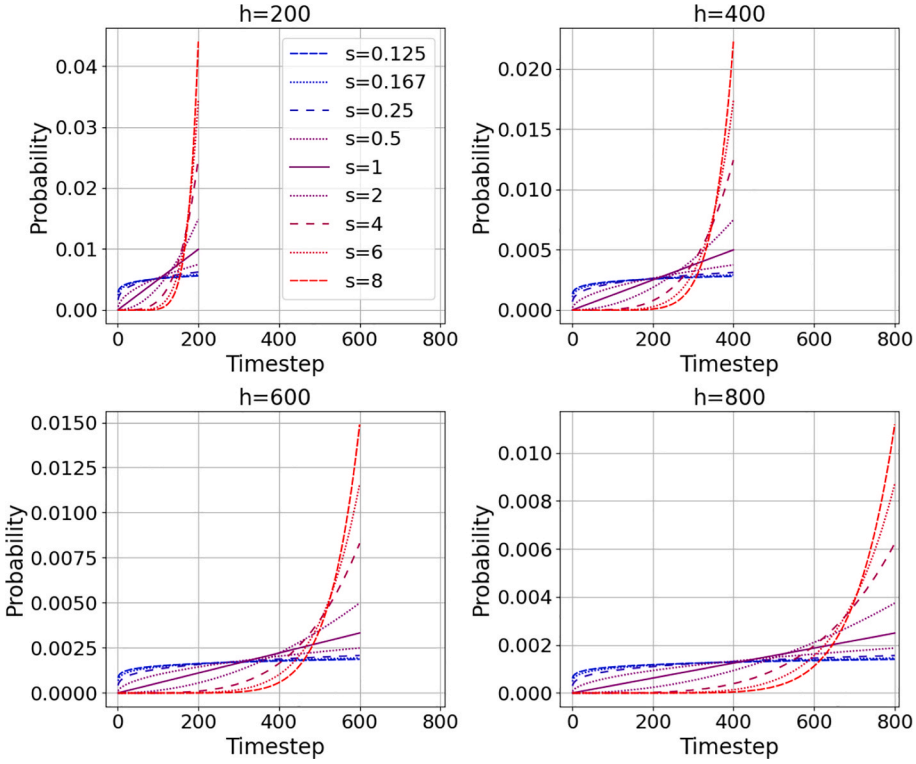


Fig. 6. Relationship between hysteresis (h) and the s exponent in Eq. 3. The rescaled probabilities lead to the cumulative probability of a saccade being exactly 1.0 at timestep h.

over all cycles. We then set the Cross value at cycle c to $\sigma_{max} - \sigma_c$.

Next, we apply the LCR. At each time step, t, we convert activations (and the pseudo 'activation' of the Cross) to *response strengths* for the value of each item, v_i:

$$r_i(t) = e^{kv_i(t)} - 1 \tag{4}$$

Note that subtracting 1 ensures that input values of 0 result in response

probabilities of 0 (another option is to skip this step and change 0 values to small positive values, e.g., 0.0001). The response strength $r_i(t)$ is a non-linear transformation of the base input value $v_i(t)$. The parameter k amplifies differences between base values (boosting high values and squashing low values). Larger values of k increase this amplification.

To obtain response probabilities that sum to 1.0 from the LCR, we simply normalize response strengths at each time step t (note that the LCR with the k parameter is algebraically equivalent to the *softmax* function with a *temperature* parameter).

$$P_i(t) = \frac{r_i(t)}{\sum_j r_j(t)} \tag{5}$$

Finally, we sample from the distribution $\{L_i(t)\}$ to select the fixation

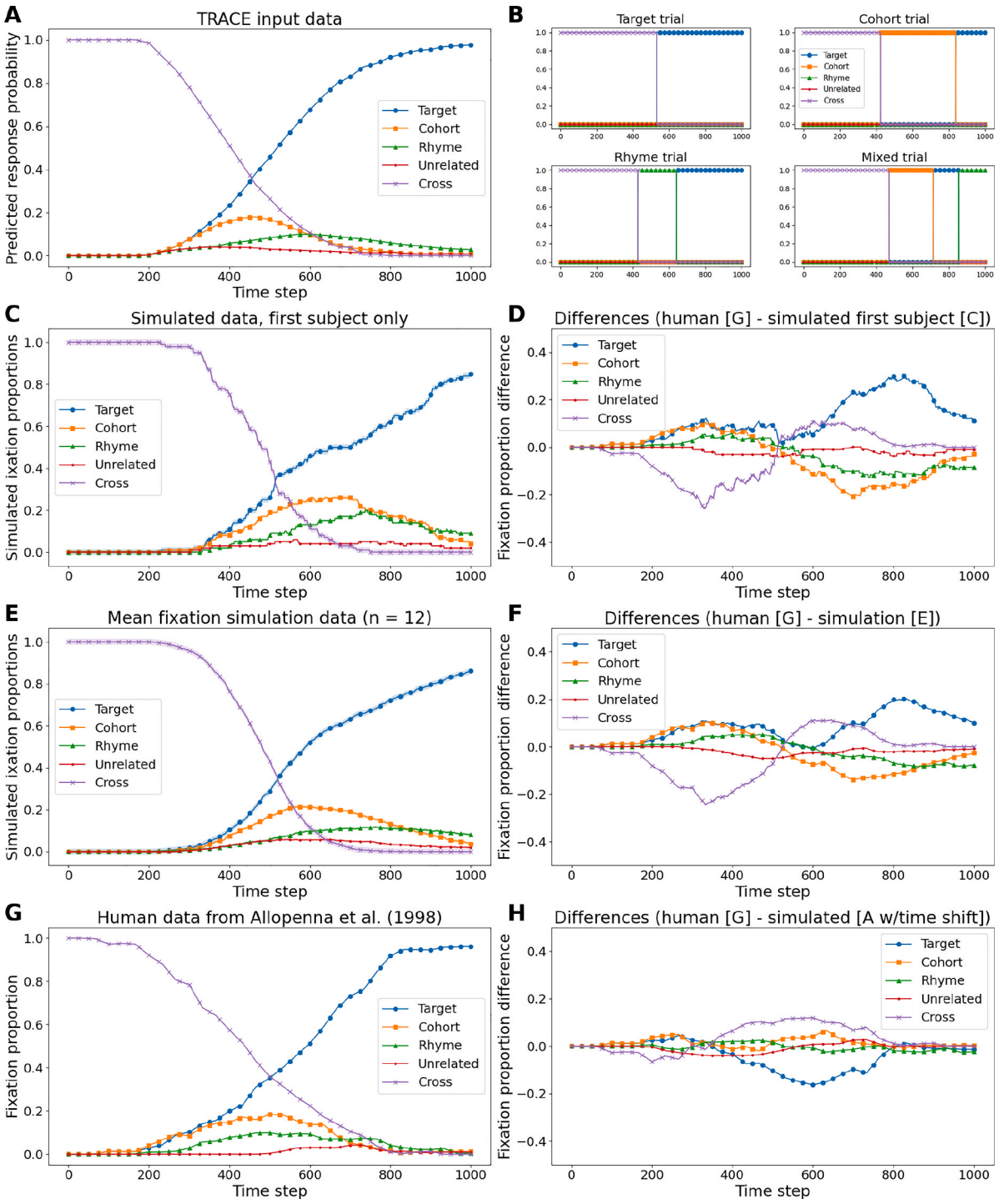


Fig. 7. A: Input to the fixation generation simulator in Simulation 1a: the predicted fixation proportions calculated by Allopenna et al. (1998). B: Examples of single trials that included only target fixations, cohort and target fixations, target and cohort fixations, and 'mixed' trials with all 3 (selected randomly from trials from the first simulated subject). C: Mean results from the first simulated subject's 100 trials (averaging over 100 trials like those in B). D: Differences between the simulated data in C and the human data shown in G. E: Mean fixations aggregated over 100 trials for each of 12 simulated subjects. F: Differences between human data and the simulated mean (G - E). G: Human data from Allopenna et al. H: Differences between human data (G) and predicted response proportions calculated by Allopenna et al. (shown in panel A). Hysteresis was set to 400 and s was set to 2.

target *when a fixation decision is triggered*, as described above. A random draw based on the probabilities $L_i(t)$ determines the fixation target (this means that if the LCR probabilities are 0.6, 0.1, 0.1, 0.1, 0.1, there is a 60% chance that the first item will be selected, and a 10% chance for each of the other items). Since the set of possible targets always includes all 5 possible items, the decision can be for the 'new' fixation target be the same as the previous fixation target.

4. Simulating fixations

Here, we will compare two bases for generating fixations. First, we will use predicted fixation proportions from Allopenna et al. (1998) (plotted in Fig. 1C) as the basis for generating fixations, and assess whether resulting mean predicted fixation proportions 'recover' the underlying predicted fixation proportions. Second, we will examine a much simpler approach that starts from unshifted TRACE activations and requires fewer free and fixed parameters.

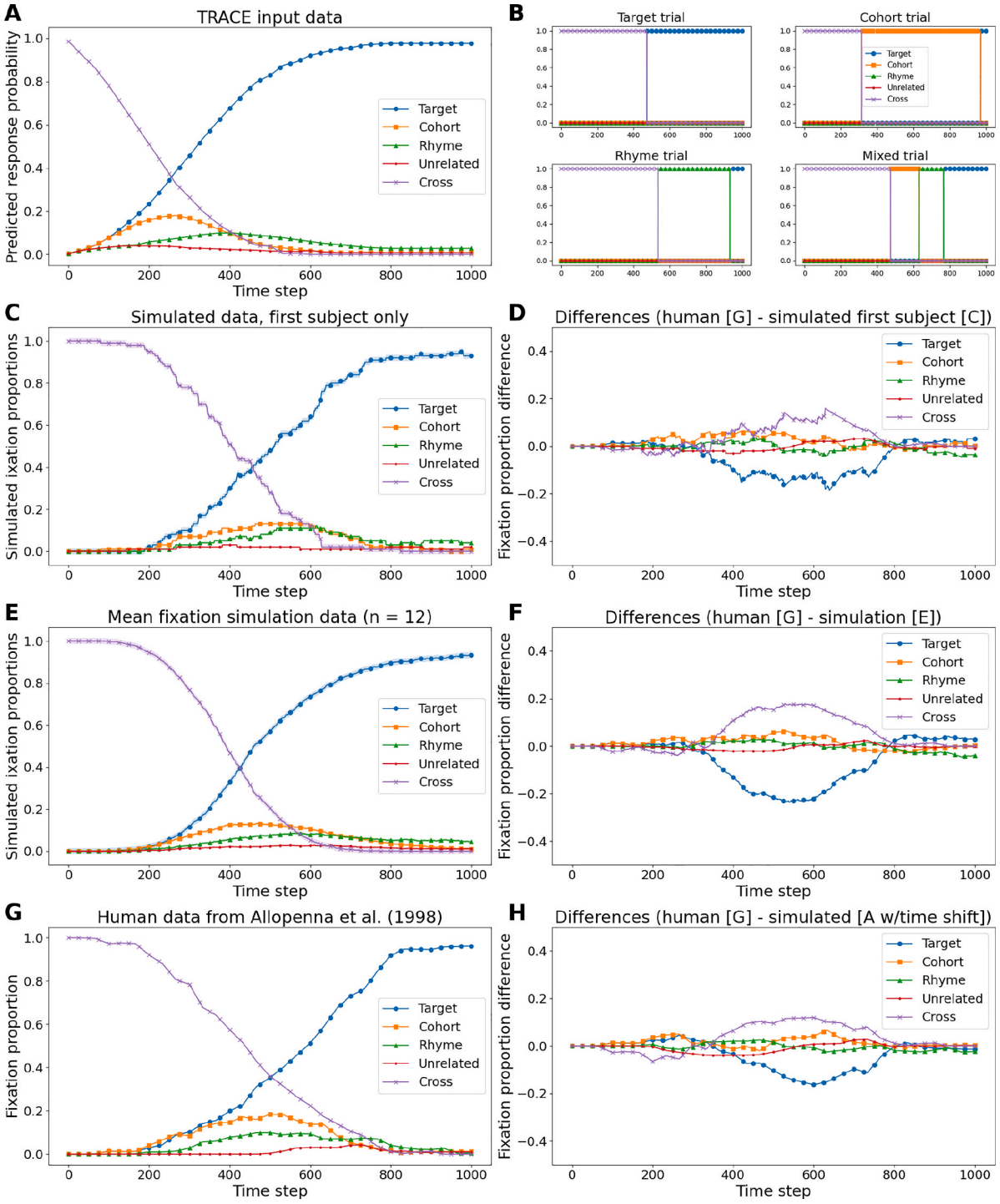


Fig. 8. A: Input to the fixation generation simulator in Simulation 1b: the predicted fixation proportions calculated by Allopenna et al. (1998), but with the time shift (of 200 ms) removed. See the caption of Fig. 7 for details of panels here. Note that panel H shows differences between Fig. 7A and panel G, just as in Fig. 7.

4.1. Simulation 1: Using pre-calculated response probabilities

4.1.1. Simulation 1a

Fig. 7 shows the results of the first batch of simulations. These were conducted using the predicted response proportions calculated by Allopenna et al. I simulated 100 trials for 12 subjects. On a given trial, fixation proportions to different items were 1 or 0 at any time step. At time 0, I stipulate that the subject is fixating the cross. Then at each subsequent time step, the process described above (with parameters *h* and *s*) is used to determine whether a fixation decision is triggered. When a fixation decision is triggered, I sample from the probabilities in Fig. 7A, and the decision can either be to continue fixating the current target or shift to another one. A little trial-and-error exploration revealed robust results with *h* set to 400 and *s* set to 2. Changing *h* to be less than ~350 or greater than ~450 leads to qualitatively different results (lower values lead to more abrupt changes and higher peaks for cohorts and rhymes; higher values lead to later fixation changes and lower cohort and rhyme peaks).

Our key interest is panel E, showing the mean from 100 simulated trials for 12 simulated subjects (though panel C gives a sense of how individual subject simulations might vary). We can observe qualitatively that the patterns in E are quite similar to those in A (predicted response proportions) and G (human VWP data). Panel F shows the differences between human data and the simulation data (G - E). Panel H shows the differences between the Allopenna et al. calculated predicted fixation proportions and the human data. We see somewhat larger discrepancies in F than H. In particular, cross proportions are too high in the early time course, while target proportions are too low in the later time course.

We can also see that in general, the patterns in E are shifted later in time compared to panels A and G. This follows from the hysteresis parameter inducing a lag between the input values and fixations. Recall that Allopenna et al. built in a time shift in their predicted fixation proportions, shifting the underlying activations such that the target proportion would become greater than zero at 200 ms. Perhaps we could improve the results if we only used one of these time-shifting parameters. We need hysteresis to simulate saccade generation, so this raises the possibility of not shifting the TRACE predictions.

4.1.2. Simulation 1b

In this simulation, we follow the same procedure as for Simulation 1a, except we 'unshift' the Allopenna et al. predicted fixation proportions, as can be seen in Fig. 8A. (Note that to then generate data for the later time course, we need to pad the activation data to compensate for the data removed at the beginning; to do so, I repeat the final values to the end of the time window.) We in fact find improvement compared to Simulation 1a. Comparing panels F of Figs. 7 and 8, we see similar lower total error in 8F, with a trade-off between better cohort and rhyme predictions and slightly worse cross and target predictions (as can be seen in 8E, the point where the cross and target cross over is a bit early).

4.1.3. Discussion

The results of Simulations 1a and 1b show that we can readily generate a plausible pattern of predicted fixation proportions by applying a simple fixation generator to the predicted fixation proportions calculated by Allopenna et al. (1998). Simulation 1b shows that we can dispense with 1 of their parameters: the temporal shift, which is no longer (or at least less) needed when we simulate fixations with a hysteresis parameter. Could we improve the fit with exhaustive parameter space exploration or even parameter optimization? Almost certainly. However, it is questionable whether this would yield substantial payoff in terms of advancing our understanding, especially since the fixation generation model is tremendously over-simplified.

Arguably, improving the details of the simplistic fixation generation component of the model would also have limited utility. What is crucial is that we see that we can robustly simulate human VWP data with two modular components: a speech processing component (TRACE in this case) and a decision process to link lexical activations to fixation behaviors. We also see, crucially, that there is very little difference between calculating response probabilities vs. simulating individual fixations. However, could this simply be due to using calculated predicted fixation proportions as the input to the fixation model? We will explore this possibility with even fewer parameters in Simulation 2.

Another crucial consideration is that the Allopenna et al. predicted fixation proportions are not meant as a basis for simulating individual trials. They are meant to describe the *output* that would result from a decision process applied to lexical activations. This is also a fundamental motivation for Simulation 2, where we will simulate fixations directly from model activations.

4.2. Simulation 2

Recall that Allopenna et al. included four free parameters when they calculated their predicted fixation proportions from TRACE activations (*k*, a scaling factor in the LCR; *x*, a steepness parameter for their 'dynamic' *k*; Δt, a scaling factor based on the ratio of maximum *activation* value at a time step relative to the maximum activation at any time step; and *Temporal shift*, how far TRACE activations were shifted later in time). In Simulation 1b, we dropped the *temporal shift* because hysteresis in the fixation generation component also induces a 'rightward' temporal shift, raising the possibility that a shift parameter is unnecessary. Let's now ask whether we can drop any of the other parameters.

First, let's drop as many parameters as possible: $x, \Delta t$, and even the LCR and its *k* parameter. Doing this implies that we will use the raw TRACE activations – which we will, but with a minor modification. Rather than using raw activations, we will rescale them so that the maximum activation value (0.59) is transformed to 1.0 and the minimum value is (still) 0.0 (without normalization; this is purely for numerical convenience). We will again compensate for the time steps removed at the beginning of the TRACE data by repeating the final activation values to the end of the time window.

For the fixation generation simulations, we will keep *h* set to 400 and *s* to 2. We will still fill in cross values as we did in Simulation 1, and again simply normalize the cross and activation values so they sum to 1 at every time step (for the fixation generation outputs, not the underlying activations; as noted above, these have been rescaled but not normalized). The results are shown in Fig. 9. The key results are in panels E and F. Panel E shows the mean time course based on 12 simulated subjects with 100 trials each. With only the *h* and *s* fixation parameters, and no decision parameters, we observe a predicted time course of phonological competition that is remarkably similar to human VWP data. The differences between panel E and panel A represent the impact of the fixation sampling procedure on the underlying activations.

There are two key discrepancies when we compare the simulated time course (E) to the human time course (G, differences in F). First, the simulated patterns are shifted earlier compared to the human data, leading to greater target proportions and lower cross proportions in the middle time course. This might be addressable with a simple temporal shift, perhaps motivated by a proposal that there is an additional delay for initial fixations (perhaps as participants begin mapping weak lexical activations to pictures).

Second, cohort and rhyme proportions are overestimated in the later time course. The fact that this is the case for cohorts even though in the underlying activations, cohort activation approaches 0 by 600 ms shows how sluggish the *h* and *s* parameters make the link between activations and fixations. For rhymes, this might be mitigated slightly if we applied the LCR, but only slightly, as we can see that the target and rhyme activations change only slightly in the late time course. Mitigating this significantly might require something like the dynamic *k* approach used by Allopenna et al. (if we used the LCR and could make *k* stronger in the later time course, this would squash rhyme proportions and boost target proportions). Whether such modifications are warranted or desirable is a question I defer for future research.

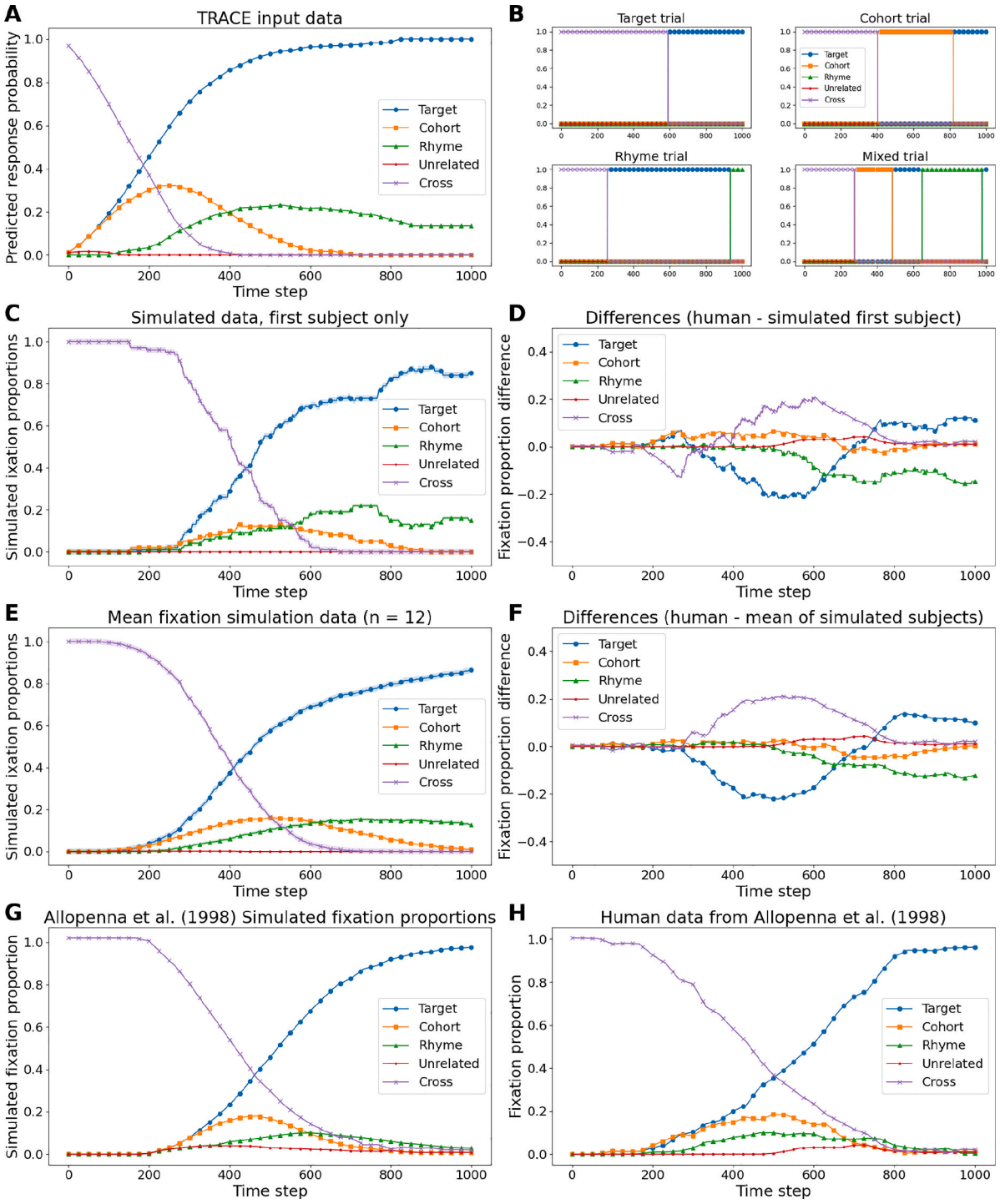


Fig. 9. Results of Simulation 2, with minimal parameters. See text for detail.

4.2.1. *Discussion*

We observed here that we can drop all decision-level parameters from previous approaches and transform activation inputs to excellent fixation proportion predictions using just a fixation generation component. This implements a trial-level decision mechanism that, with just 2 parameters, provides a linking hypothesis that maps activation patterns (which already provide the key phonological competition time course) to highly accurate fixation proportion predictions over time. While I noted ways we could potentially improve fit, as I noted in Simulation 1, there is little apparent utility in doing so. That said, the simulation code is freely available, and others may wish to explore these details.

5. General Discussion

Simulations 1a and 1b demonstrate that a simple, 2-parameter fixation generation model driven by predicted fixation proportions derived from the linking hypothesis of Allopenna et al. (1998) robustly simulates VWP fixation proportions. On the one hand, this was an exercise in demonstrating the nearly obvious: probabilistically sampling repeatedly from categories of quantities that vary over time recovers something close to the underlying quantities. On the other, we also discover something new: the assumption that there is a refractory period between saccades of probabilistic duration allows us to dispense with the

'temporal shift' parameter used by Allopenna et al. However, sampling from predicted response proportions is somewhat strange; those proportions represent the prediction of the outcome of a decision process.

Thus, in Simulation 2, I used normalized activations to drive the fixation model. This allowed me to drop all decision parameters, leaving only the parameters of the fixation model. The mildly surprising result was that this yields predicted fixation proportions that are only slightly less precise than the predictions Allopenna et al. calculated via their linking hypothesis, without requiring the use of the Luce choice rule. Virtually the simplest possible decision process is implemented, and the outcome actually closely resembles the LCR-based predictions from Allopenna et al. This makes sense if we consider the LCR a descriptive mathematical model of principles in biological decision making rather than an algorithmic process model.

The correspondence between simulated fixations and fixation proportions over time in Simulation 1 does not somehow validate the Allopenna et al. linking hypothesis; no such validation was required. Its validity was already demonstrated by Allopenna et al. It seems that the critique of Norris (2005) was based on a misunderstanding of the relation between predicted *fixation* probabilities (which map to fixation proportions over time) and perhaps predicted *saccade* probabilities, which were not the target of the linking hypothesis.

That said, simulating trial-level saccades and fixations rather than using the LCR to derive a linking hypothesis has been a useful exercise. It should assuage any lingering doubts about whether the LCR-based approach corresponds closely to a trial-level approach. More usefully, Simulation 2 shows that a simple trial-level fixation decision module lacking the parameters of the LCR results in predicted fixation proportions that closely correspond to LCR predictions and human VWP data. An intriguing potential extension of this approach might be to examine individual differences in VWP data, as discussed earlier. Individuals who appear to show weak and/or delayed competition may have atypical underlying lexical activations – or they may have a sluggish fixation decision process.

To get a sense of this potential, consider Fig. 10. If we lower h to 300, the simulation appears to show greater cohort competition. If we increase h to 500 or 800, we substantially diminish apparent cohort competition. In each case, the underlying lexical activations are identical; only the sampling parameters for fixation decisions have changed. Thus, this approach opens the possibility of distinguishing 'output'-level bases vs. lexical activation bases for individual differences (cf. Mirman et al., 2011, who found that the LCR response selectivity parameter provided the most parsimonious basis for simulating differences in cohort competition between Broca's and Wernicke's aphasia patients).

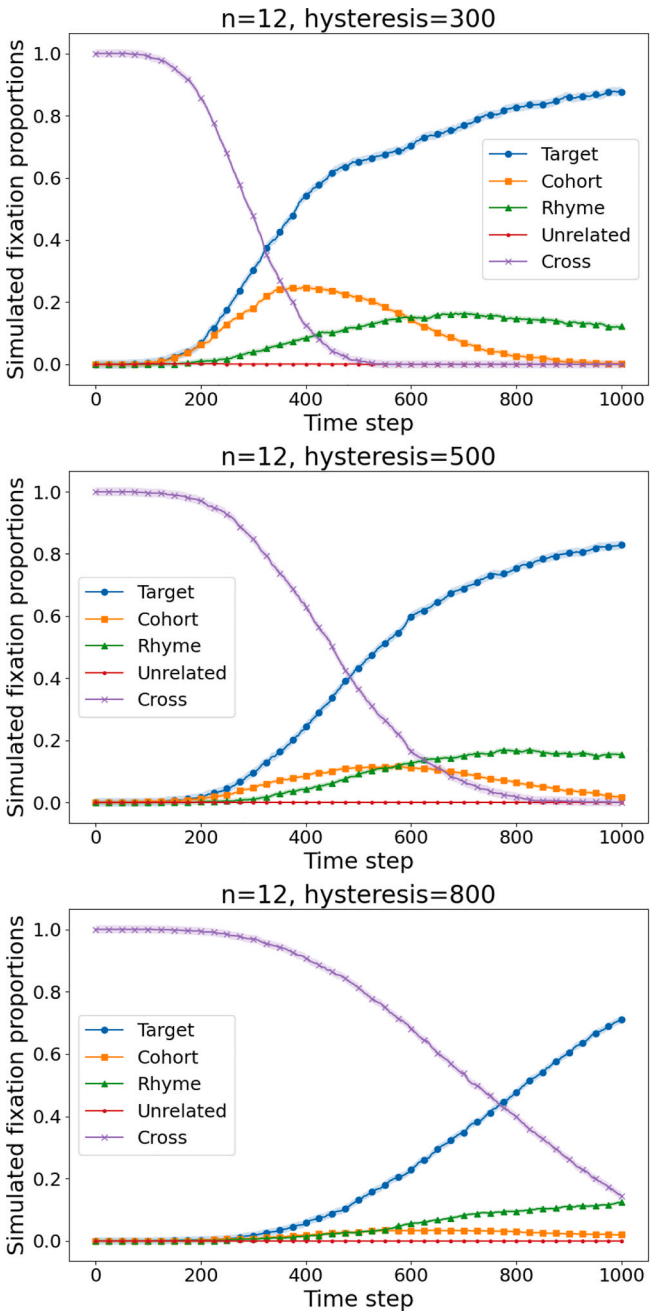


Fig. 10. Impact of hysteresis on apparent competition.

6. Conclusions

In conclusion, the VWP critique that motivated this paper (Norris, 2005) appears to have been itself motivated by a misinterpretation of VWP data (possibly due to Allopenna et al. describing fixation proportions as fixation *probabilities*, or an interpretation of predicted fixations as predicted saccades). Nonetheless, it inspired the detailed comparisons presented here between VWP and other psychological data (such as lexical decision data), and the relative merits of focusing on central tendencies, specific items, and trial-level behavior. It also inspired the fixation-generation simulations presented here, which serve as a possible foundation for deeper investigations into trial-level behavior and individual differences.

Such an effort could also provide a foundation for explorations of the proposal by Spivey (2025) that there is interaction between fixations and language processing (e.g., the object one is fixating will influence lexical processing via feedback to lexical processing). However, if an investigation is not concerned with individual differences that might emerge from trial-level behavior, or Spivey's deep interaction hypothesis, there is arguably no need to simulate trial-level behavior. As I have discussed, central tendencies are appropriate and standard targets for computational model simulations.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Allopenna, P.D., 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. https://doi.org/10.1006/jmla.1997.2558.

Balota, D.A., 1990. The role of meaning in word recognition. In: Balota, D.A., Flores d'Arcais, G.B., Rayner, K. (Eds.), Comprehension processes in reading. Lawrence Erlbaum Associates, Hillsdale, N.J, pp. 9–32.

Cooper, R.M., 1974. The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. Cognitive Psychology 6 (1), 84–107. https://doi.org/10.1016/0010-0285(74)90005-X.

Dahan, D., Magnuson, J.S., Tanenhaus, M.K., 2001. Time course of frequency effects in spoken-word recognition: Evidence from eye movements. Cognitive Psychology 42 (4), 317–367. https://doi.org/10.1006/cogp.2001.0750.

Dahan, D., Magnuson, J.S., Tanenhaus, M.K., Hogan, E.M., 2001. Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. Language and Cognitive Processes 16 (5), 507–534. https://doi.org/10.1080/01690960143000074.

Elman, J.L., 1990. Finding structure in time. Cognitive Science 14 (2), 179–211. https://doi.org/10.1207/s15516709cog1402_1.

Elman, J.L., 1991. Distributed representations, simple recurrent networks, and grammatical structure. Machine Learning 7 (2), 195–225. https://doi.org/10.1007/BF00114844.

Grainger, J., Jacobs, A.M., 1996. Orthographic processing in visual word recognition: A multiple read-out model. Psychological Review 103 (3), 518–565. https://doi.org/10.1037/0033-295X.103.3.518.

Hannagan, T., Magnuson, J.S., Grainger, J., 2013. Spoken word recognition without a TRACE. Frontiers in Psychology 4. https://doi.org/10.3389/fpsyg.2013.00563.

Jacobs, A.M., Grainger, J., 1994. Models of visual word recognition: Sampling the state of the art. Journal of Experimental Psychology: Human Perception and Performance 20 (6), 1311–1334. https://doi.org/10.1037/0096-1523.20.6.1311.

Lee, C.L., Estes, W.K., 1977. Order and position in primary memory for letter strings. Journal of Verbal Learning and Verbal Behavior 16 (4), 395–418. https://doi.org/10.1016/S0022-5371(77)80036-4.

Lee, C.L., Estes, W.K., 1981. Item and order information in short-term memory: Evidence for multilevel perturbation processes. Journal of Experimental Psychology: Human Learning and Memory 7 (3), 149–169. https://doi.org/10.1037/0278-7393.7.3.149.

Li, M.Y.C., Braze, D., Kukona, A., Johns, C.L., Tabor, W., Van Dyke, J.A., Magnuson, J.S., 2019. Individual differences in subphonemic sensitivity and phonological skills. Journal of Memory and Language 107, 195–215. https://doi.org/10.1016/j.jml.2019.03.008.

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. Ear and Hearing, 19 (1), 1. Retrieved August 9, 2024, from https://journals.lww.com/ear-hearing/fulltext/1998/02000/recognizing_spoken_words__the_neighborhood.1.aspx?casa_token=t-5S474VUUoAAAAA:YgW-cLQICALYWKedY1-arRiBTVRQZsz6CmT_lKqnp1F_XkSEI6QMpoaPdfjM2SPmlwEZoZPfi_Zq6KR_CsGmxhE.

Luce, R.D., 1959. On the possible psychophysical laws. Psychological Review 66 (2), 81–95. https://doi.org/10.1037/h0043178.

Magnuson, J.S., 2019. Fixations in the visual world paradigm: Where, when, why? Journal of Cultural Cognitive Science 3 (2), 113–139. https://doi.org/10.1007/s41809-019-00035-3.

Magnuson, J.S., 2024. Linking lexical decision to computational models. https://doi.org/10.6084/M9.FIGSHARE.27273327.V1.

Magnuson, J.S., 2024. Linking the visual world paradigm to computational models. https://doi.org/10.6084/M9.FIGSHARE.27273381.V1.

Magnuson, J.S., 2024. Overlap in psycholinguistic measures. https://doi.org/10.6084/M9.FIGSHARE.27273309.V1.

Magnuson, J.S., Dixon, J.A., Tanenhaus, M.K., Aslin, R.N., 2007. The dynamics of lexical competition during spoken word recognition. Cognitive Science 31 (1), 133–156. https://doi.org/10.1080/03640210709336987.

Magnuson, J.S., Mirman, D., Harris, H.D., 2012. Computational models of spoken word recognition. In: Spivey, M., McRae, K., Joanisse, M. (Eds.), The Cambridge Handbook of Psycholinguistics. Cambridge University Press, pp. 76–103.

Magnuson, J.S., Tanenhaus, M.K., Aslin, R.N., Dahan, D., 2003. The time course of spoken word learning and recognition: Studies with artificial lexicons. Journal of Experimental Psychology: General 132 (2), 202–227. https://doi.org/10.1037/0096-3445.132.2.202.

Magnuson, J.S., You, H., Luthra, S., Li, M., Nam, H., Escabí, M., Rueckl, J.G., 2020. EARSHOT: A minimal neural network model of incremental human speech recognition. Cognitive Science 44 (4), e12823. https://doi.org/10.1111/cogs.12823.

Marslen-Wilson, W., Tyler, L.K., 1980. The temporal structure of spoken language understanding. Cognition 8 (1), 1–71. https://doi.org/10.1016/0010-0277(80)90015-3.

Marslen-Wilson, W., Warren, P., 1994. Levels of perceptual representation and process in lexical access: Words, phonemes, and features. Psychological Review 101 (4), 653–675. https://doi.org/10.1037/0033-295X.101.4.653.

Marslen-Wilson, W., Zwitserlood, P., 1989. Accessing spoken words: The importance of word onsets. Journal of Experimental Psychology: Human Perception and Performance 15 (3), 576–585. https://doi.org/10.1037/0096-1523.15.3.576.

Marslen-Wilson, W.D., Welsh, A., 1978. Processing interactions and lexical access during word recognition in continuous speech. Cognitive Psychology 10 (1), 29–63. https://doi.org/10.1016/0010-0285(78)90018-X.

Marslen-Wilson, W.D., 1987. Functional parallelism in spoken word-recognition. Cognition 25 (1), 71–102. https://doi.org/10.1016/0010-0277(87)90005-9.

Matin, E., Shao, K.C., Boff, K.R., 1993. Saccadic overhead: Information-processing time with and without saccades. Perception & Psychophysics 53 (4), 372–380. https://doi.org/10.3758/BF03206780.

McClelland, J.L., Elman, J.L., 1986. The TRACE model of speech perception. Cognitive Psychology 18 (1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0.

McMurray, B., 2023. 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. Psychonomic Bul letin & Review 30 (1), 102–146. https://doi.org/10.3758/s13423-022-02143-8.

McQueen, J.M., Norris, D., Cutler, A., 1999. Lexical influence in phonetic decision making: Evidence from subcategorical mismatches. Journal of Experimental Psychology: Human Perception and Performance 25 (5), 1363–1389. https://doi.org/10.1037/0096-1523.25.5.1363.

Mirman, D., 2014. Growth curve analysis and visualization using r. https://doi.org/10.1201/9781315373218.

Mirman, D., Dixon, J.A., Magnuson, J.S., 2008. Statistical and computational models of the visual world paradigm: Growth curves and individual differences. Journal of Memory and Language 59 (4), 475–494. https://doi.org/10.1016/j.jml.2007.11.006.

Mirman, D., Yee, E., Blumstein, S.E., Magnuson, J.S., 2011. Theories of spoken word recognition deficits in aphasia: Evidence from eye-tracking and computational modeling. Brain and Language 117 (2), 53–68. https://doi.org/10.1016/j.bandl.2011.01.004.

Norris, D. (2005). How do computational models help us develop better theories? In Twenty-first century psycholinguistics: Four cornerstones (pp. 331–346). doi: 10.4324/9781315084503-24.

Norris, D., McQueen, J.M., 2008. Shortlist b: A bayesian model of continuous speech recognition. Psychological Review 115 (2), 357–395. https://doi.org/10.1037/0033-295X.115.2.357.

Norris, D., McQueen, J.M., Cutler, A., 2000. Merging information in speech recognition: Feedback is never necessary. Behavioral and Brain Sciences 23 (3), 299–325. https://doi.org/10.1017/S0140525X00003241.

Oleson, J.J., Cavanaugh, J.E., McMurray, B., Brown, G., 2017. Detecting time-specific differences between temporal nonlinear curves: Analyzing data from the visual world paradigm. Statistical Methods in Medical Research 26 (6), 2708–2725. https://doi.org/10.1177/0962280215607411.

Roberts, S., Pashler, H., 2000. How persuasive is a good fit? a comment on theory testing. Psychological Review 107 (2), 358–367. https://doi.org/10.1037/0033-295X.107.2.358.

Roberts, S., Pashler, H., 2002. Reply to rodgers and rowe (2002). Psychological Review 109 (3), 605–607. https://doi.org/10.1037/0033-295X.109.3.605.

Smith, A.C., Monaghan, P., Huettig, F., 2017. The multimodal nature of spoken word processing in the visual world: Testing the predictions of alternative models of multimodal integration. Journal of Memory and Language 93, 276–303. https://doi.org/10.1016/j.jml.2016.08.005.

Spivey, M., 2006. The Continuity of Mind. Oxford University Press, New York.

Spivey, M.J., 2025. A linking hypothesis for eyetracking and mousetracking in the visual world paradigm. Brain Research, 1851. https://doi.org/10.1016/j.brainres.2025.149477.

Spivey-Knowlton, M.J., 1996. Integration of visual and linguistic information: Human data and model simulations. University of Rochester.

Tanenhaus, M.K., Magnuson, J.S., Dahan, D., Chambers, C., 2000. Eye movements and lexical access in spoken-language comprehension: Evaluating a linking hypothesis between fixations and linguistic processing. Journal of Psycholinguistic Research 29 (6), 557–580. https://doi.org/10.1023/A:1026464108329.

Tanenhaus, M.K., Magnuson, J.S., McMurray, B., Aslin, R.N., 2000. No compelling evidence against feedback in spoken word recognition. Behavioral and Brain Sciences 23 (3), 348–349. https://doi.org/10.1017/S0140525X0048324X.

Tanenhaus, M.K., Spivey-Knowlton, M.J., 1996. Eye-tracking. Language and Cognitive Processes 11 (6), 583–588. https://doi.org/10.1080/016909696386971.

Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., Sedivy, J.C., 1995. Integration of visual and linguistic information in spoken language comprehension. Science 268 (5217), 1632–1634. https://doi.org/10.1126/science.7777863.