WALD-EM: Wald Accumulation for Locations and Durations of Eye Movements

Kucharský, Š.; van Renswoude, D.; Raijmakers, M.; Visser, I.

DOI
10.31234/osf.io/2s48r
10.1037/rev0000292

Publication date
2021

Document Version
Final published version

Published in
Psychological Review

License
Article 25fa Dutch Copyright Act (https://www.openaccess.nl/en/in-the-netherlands/you-share-we-take-care)

Link to publication

Citation for published version (APA):
Kucharský, Š., van Renswoude, D., Raijmakers, M., & Visser, I. (2021). WALD-EM: Wald Accumulation for Locations and Durations of Eye Movements. *Psychological Review*, *128*(4), 667-689. https://doi.org/10.31234/osf.io/2s48r, https://doi.org/10.1037/rev0000292

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Review

© 2021 American Psychological Association
ISSN: 0033-295X

2021, Vol. 128, No. 4, 667–689
https://doi.org/10.1037/rev0000292

WALD-EM: Wald Accumulation for Locations and Durations of Eye Movements

Šimon Kucharský[1], Daan van Renswoude[2], Maartje Raijmakers[1, 3], and Ingmar Visser[1, 4]
[1] Faculty of Social and Behavioural Sciences, Department of Psychology, University of Amsterdam
[2] Center of Linguistics, Faculty of Humanities, Leiden University
[3] Faculty of Behavioral and Movement Sciences, Department of Educational Studies and Learn, Free University Amsterdam
[4] Amsterdam Brain and Cognition, University of Amsterdam

Describing, analyzing, and explaining patterns in eye movement behavior is crucial for understanding visual perception. Further, eye movements are increasingly used in informing cognitive process models. In this article, we start by reviewing basic characteristics and desiderata for models of eye movements. Specifically, we argue that there is a need for models combining spatial and temporal aspects of eye-tracking data (i.e., fixation durations and fixation locations), that formal models derived from concrete theoretical assumptions are needed to inform our empirical research, and custom statistical models are useful for detecting specific empirical phenomena that are to be explained by said theory. In this article, we develop a conceptual model of eye movements, or specifically, fixation durations and fixation locations, and from it derive a formal statistical model—meeting our goal of crafting a model useful in both the theoretical and empirical research cycle. We demonstrate the use of the model on an example of infant natural scene viewing, to show that the model is able to explain different features of the eye movement data, and to showcase how to identify that the model needs to be adapted if it does not agree with the data. We conclude with discussion of potential future avenues for formal eye movement models.

Keywords: eye movements, eye-tracking, fixation durations, fixation locations, evidence accumulation

Supplemental materials: https://doi.org/10.1037/rev0000292.supp

As only a relatively small region on the retina provides the highest detail of the visual input, the human visual system heavily relies on the ability to control the gaze and movement of the eye over a stimulus (Duchowski, 2007). Much of the current research intends to determine the mechanisms and factors[1] that guide visual attention through fixations and saccades, that is, periods of fixing the visual input relatively steady on the retina and periods of abrupt movements, respectively, as understanding these mechanisms provides insights into visual and attentional control and their impact on perception. Additionally, studying eye movements is not only essential for understanding perception and attentional control but can also inform variety of other topics, such as the study of higher cognitive processes like decision rules in economic games (Polonio et al., 2015), strategic differences in analogical reasoning tasks (Hayes et al., 2015; Kucharský et al., 2020), or individual assessment (Chen et al., 2014), to name a few.

Previous research distinguishes the mechanisms and factors that guide visual attention into three groups (Itti & Borji, 2014; Schütt et al., 2017; Tatler & Vincent, 2008). These groups can be roughly described as bottom-up, top-down, and systematic tendencies. The bottom-up factors include features of the visual environment, such as distribution of colors and contrast across the visual field, etc. Many of the so-called saliency models aim to determine and detect these features (Itti & Koch, 2001; Tatler et al., 2011; Xu et al., 2014). The top-down factors and mechanisms include characteristics and states of the observer, such as their motivation, purpose, task, (background) knowledge, or individual differences (De Haas et al., 2019). The third group includes factors that are neither purely bottom-up (i.e., not necessarily tied to features in the environment) nor top-down (i.e., not necessarily unique to states or characteristics

This article was published Online First May 27, 2021.
Šimon Kucharský https://orcid.org/0000-0003-4192-1140
Maartje Raijmakers https://orcid.org/0000-0003-1843-6462
Ingmar Visser https://orcid.org/0000-0003-3855-2778
Maartje Raijmakers and Ingmar Visser share the final authorship.

The data that were reanalyzed in this article were presented in van Renswoudeet al. (in prep), and in the PhD dissertation and defense of Daan van Renswoude. The modeling work in this manuscript was presented at a Code-P meeting at University of Amsterdam, 24th November 2020. The data and code associated with this article are publicly available at https://github.com/Kucharssim/WALD-EM and the draft of the article is available as a preprint at https://psyarxiv.com/2s48r/.

The data, code and other materials are publicly available at the project's public repository: https://github.com/Kucharssim/WALD-EM.

The analyses in this manuscript are purely exploratory; all data were collected prior to the analysis. None of the analyses were preregistered. All modeling decisions were made after we have collected the data, but we did in fact split the data set in a training and test set before any modeling exercises.

Šimon Kucharský was supported by the NWO (Nederlandse Organisatie voor Wetenschappelijk Onderzoek) grant no. 406.10.559. We would like to thank Andrew Heathcote and an anonymous reviewer for their constructive and thought-provoking reviews of the manuscript.

Correspondence concerning this article should be addressed to Šimon Kucharský, Faculty of Social and Behavioural Sciences, Department of Psychology, University of Amsterdam, Nieuwe Achtergracht 129-B, Amsterdam 1018 WS, the Netherlands. Email: s.kucharsky@uva.nl

[1] Throughout the article, we use the term "factor" as "a circumstance, fact, or influence that contributes to a result" without having a specific functional form of the relationship in mind.

of observers), but rather experimentally observed phenomena (Tatler & Vincent, 2008). Systematic tendencies are believed to be relatively stable across stimuli, participants, and tasks, such as fixation biases (e.g., central bias; Tatler, 2007; Tseng et al., 2009; van Renswoude et al., 2019) or saccadic biases (e.g., horizontal and leftward bias; Foulsham et al., 2013, 2018; Le Meur & Liu, 2015; van Renswoude et al., 2016).

Apart from experimental work establishing individual factors that influence gaze behavior, important aspect of understanding the mechanism behind the observed behavior is proposing theoretical and statistical models that are able to describe, explain, or predict empirical data and observed phenomena. There are many models with varying levels of abstraction, theoretical substance, the phenomena they aim to explain, and the type and level of data they are able to explain (Le Meur & Liu, 2015; Malem-Shinitski et al., 2020; Nuthmann, 2017; Reichle & Sheridan, 2015; Schütt et al., 2017; Schwetlick et al., 2019; Tatler et al., 2017; Trukenbrod & Engbert, 2014; Zelinsky et al., 2013). In this article, we develop a new conceptual model of eye movements, and flesh it out in the form of a statistical model.

Model Requirements

Two prominent questions regarding eye movement behavior that require explanation are *when* and *where* (Findlay & Walker, 1999; Tatler et al., 2017), that is, what is the mechanism behind the *timing of saccades and fixation durations*, and what is the mechanism behind selecting *fixation locations*. Predominantly, these questions are asked separately by building models explaining either fixation durations or fixation locations (Nuthmann et al., 2010; Schütt et al., 2017; Tatler et al., 2017). However, better understanding of visual behavior is perhaps only possible when considering *where* and *when* people look simultaneously (Tatler et al., 2017). It is of interest to consider spatial and temporal phenomena in one model, as these are likely not independent of each other (e.g., Henderson et al., 2013; Nuthmann, 2017). In this article, we propose a new account of how to model eye movements both spatially and temporally in a joint framework.

One of the critical features of theory-driven models of any behavior is the ability to generate data, given its set of assumptions. This enables to assess whether a model is successful in generating phenomena that are putatively explained by said theory (Borsboom et al., 2020; Robinaugh et al., 2020), and also makes it possible to make counterfactual investigations. That is, we might use it to answer the question "according to the model, what would have happened if something would have occurred, but it did not?," which is useful for hypothesis generation and essentially more precise testing of theories underlying the models (e.g., Nuthmann et al., 2010). This is generally a useful approach that enables to check the explanatory adequacy of the underlying theory, inform us about where to look for crucial piece of evidence, and as such serving a crucial part of the theoretical cycle (Borsboom et al., 2020). Building data generative models of eye movements have a long tradition in the eye-tracking literature. In fact, the traditional approach to evaluate eye movement models typically involves simulating eye movement data from a model and comparing the synthetic data to experimentally established phenomena (Schütt et al., 2017).

Additionally to being used as generative models, formal modeling approaches are widely used in the empirical cycle as well in form of statistical models, where they play a crucial role in detecting and establishing new phenomena from the collected data (Wagenmakers et al., 2018). Thus, as dynamic models of eye movements gain importance in theoretical and experimental research, parameter estimation and model comparison are also gaining importance. This requires being able to specify a model as a statistical model (i.e., a probability distribution of the data given a set of parameters) that can be used to estimate the parameters (either using maximum likelihood or Bayesian approaches), and use the statistical machinery for assessing the uncertainty in parameter estimates and to conduct model comparisons (e.g., Malem-Shinitski et al., 2020; Schütt et al., 2017).

Detecting new phenomena is of great interest for eye movement researchers. For example, in studying phenomena such as the central bias (i.e., relative preference to focus on the center of the screen compared to other areas), there is an ongoing debate whether it can be explained away as a manifestation of bottom-up effects (such as distributions of objects on the screen) or whether it is a real systematic bias somehow ingrained in our visual system, and how to disentangle these explanations (Tatler, 2007; Tseng et al., 2009; van Renswoude et al., 2019). Having a possibility to modify the model such that it includes or excludes the central bias, would enable us to pit these explanations against each other. Through model comparison and parameter estimation, we can then assess whether and quantify to what extent these different factors come into play. Thus, it is important that a model can be modified to include, exclude, or modify the functional form of the effect of different factors or mechanisms influencing the eye movement behavior.

Furthermore, it is highly likely that eye movement characteristics will depend on individual differences, differences between different populations, or within-person differences due to development (De Haas et al., 2019). It is thus important to be able to model these differences in one coherent modeling framework by allowing to specify parameters in the model to, for example, differ between populations or as random terms in a hierarchical fashion.

Models that are possible to use both in the theoretical cycle (i.e., as formal manifestations of a theory to check that the theory explains phenomena that it set out to explain) and empirical cycle (i.e., to assess the evidence for new phenomena that need explanation) are generally difficult to develop and rare, so rare that these two purposes of scientific models are often discussed as completely separate entities (Smaldino, 2017). However, having a model that is both statistical and informed by the underlying theory often offers deeper insights into the underlying mechanisms (Borsboom et al., 2020; Rodgers, 2010), and provides additional opportunities to learn both about the model and the natural phenomena (McElreath, 2020, pp. 525–552). In cognitive psychology, such models are sometimes referred to as cognitive process models, as they describe the cognitive processes that underlie the data, and possess parameters that often have clear interpretations (Forstmann & Wagenmakers, 2015).

Outline

In this article, we propose a model that explains fixation locations and fixation durations simultaneously, and is (a) generative (i.e., can make predictions about the locations of fixations at a particular time), (b) statistical (i.e., has a proper likelihood function),

(c) modifiable (i.e., can be expanded to include different factors, including random factors), and (d) can be interpreted as a cognitive process model.

The structure of this article is as follows. In the next section, we introduce the model in conceptual terms, that is, describe the architecture of the model to highlight the core assumptions which yield the model interpretable as a cognitive process model, while abstracting from particular analytic choices. Then, we show how to derive a particular realization of the model. This will involve laying out concretely what analytic choices we made to make the model tractable. We lay out several factors that can optionally be included in the model, and apply different versions of the model to real data to answer substantive questions, thereby illustrating the model flexibility and usefulness. In the following sections, we limit ourselves mostly to the domain of free scene viewing. We believe that the proposed model could be extended or adapted further to other contexts or paradigms, but that is not the focus of the current article.

We will refer to the new model as WALD-EM, standing for "Wald accumulation of locations and durations of eye movements." Reasons for this name will become apparent in the following description of the model.

Conceptual WALD-EM Model

Our model describes eye movement data as *x* and *y* coordinates and durations of fixations, and aims to provide answers to the questions about *when* and *where* simultaneously. As such, it consists of two parts: One that corresponds to the question *when*, and one that corresponds to the question *where*. These two parts are then intertwined together to capture potential dependencies between these two.

Model for *When*

A typical human (adult) in typical situations makes on average one saccade in 200–400 msec. The distribution of fixation durations is characteristically positively skewed with typically positive relationship between the mean and a variance, much like typical distributions of response times in decision tasks (Palmer et al., 2011). Hence, it is reasonable to borrow from the response time modeling literature, that is, evidence accumulation models, such as LATER (Carpenter & Williams, 1995), Linear Ballistic Accumulation (Brown & Heathcote, 2008), or Diffusion Decision (Ratcliff & McKoon, 2008).

In our model, we represent the fixation duration as the time it takes the observer to make a decision to make a saccade. The decision process represents information uptake from a current location up to a point where the currently fixated location does not bring additional information compared to potential information sources at other locations. We assume that information uptake is a continuous-time stochastic process that rises to a threshold with some drift and noise level. The time to make the decision to make a saccade is the first passage time of this process. The simplest model for such a time is the Wald distribution with three parameters: drift (ν), decision boundary (α), and standard deviation of the noise (σ), one of which needs to be fixed for identifiability purposes (Chhikara & Folks, 1988). Apart from that the Wald distribution is a reasonable candidate as it reflects the noisy evidence accumulation process (a process that has been deemed as a neurally plausible mechanism for decision processes, Anders et al., 2016), it has previously also been shown to fit fixation durations well (Palmer et al., 2011). Figure 1 shows the mechanism that gives rise to the Wald distribution.

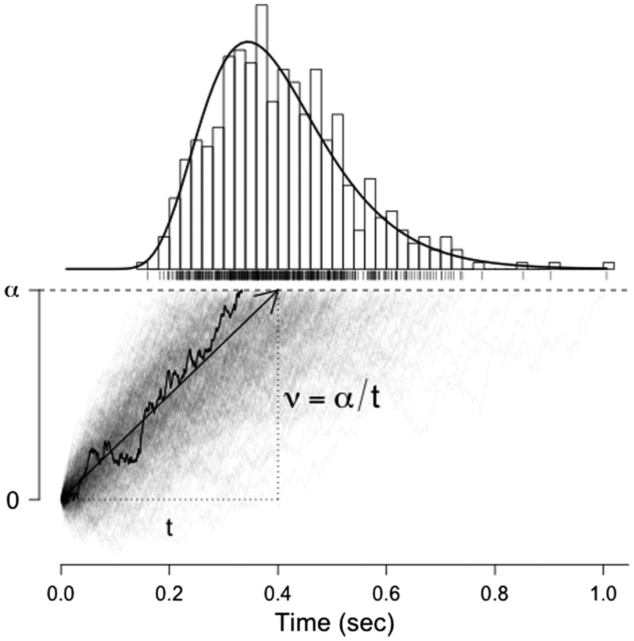
Other models contain similar data generating processes for fixation durations. For example, the LATEST model (Tatler et al., 2017) assumes that the fixation duration is also the time to make a decision to make a saccade to a new location. Our model assumes stochastic random walk accumulation, whereas LATEST assumes a linear ballistic process. Further our model assumes only one decision process at a time, whereas LATEST assumes many accumulators running in parallel. CRISP (Nuthmann et al., 2010) and ICAT (Trukenbrod & Engbert, 2014) models also rely on a stochastic random walk underlying the fixation durations. In CRISP and ICAT, however, decisions to make a saccade can be canceled by additional processes, whereas our proposal is simpler in that passing the threshold immediately triggers a saccade. Further, in ICAT and CRISP, the stochastic rise to threshold is thought of as an autonomous timer, suggesting an inherent (but stochastic) rhythmically to saccades, whereas our accumulator depends not only on internal characteristics of the observer, but their surroundings as well.

Model for *Where*

After the observer concludes that there is an advantage to move to another location, it is time to make a saccade.

Figure 1

Illustration of the Process That Results in a Wald Distribution. Evidence Starts at 0 and Accumulates as a Wiener Process With a Drift ν (displayed as Arrow) Until it Reaches a Threshold α. The Process is Inherently Noisy as Shown by 500 Different Traces Generated With the Same Parameters (Grey Lines). The First Passage Time (The Time it Takes to Trespass the Threshold α for the First Time) Results in a Wald Distribution (Displayed on Top)



Each location of the stimulus provides some amount of attraction to the observer. We call a function that maps the stimulus coordinates to that attraction an *intensity* function and denote it as: $\lambda : \mathbb{R}^2 \rightarrow \mathbb{R}_+$, and will write it as $\lambda(x, y|s)$, where s stands for the current fixation. The total amount of intensity of the whole stimulus is the integral (sum) of all the points of the stimulus: $\Lambda = \int\int \lambda(x, y|s)dxdy$. In essence, we assume that when observers decide *where* to go next, they pick a random location from a distribution proportional to this function. The function may or may not depend on the current or previous fixations, depending on whether we assume a homogeneous (static over time) or heterogeneous (evolving over time) process, and can be adjusted depending on the researcher's questions and desires.

In general, we will represent the intensity function as a combination of different factors that influence the intensity of different locations. These factors may represent different features and can be combined in different ways (see Barthelmé et al., 2013). For example, we can build the intensity function such that it combines bottom-up features of the stimulus (e.g., saliency) with systematic tendencies (e.g., central bias or horizontal bias), and so forth. Some of the factors can be thought of as representing information provided by the stimulus, assuming that locations that are rich in the information they provide will be attractive to fixate—and so will have a high intensity. However, people not always fixate on locations with a lot of information. Later, it will be important to make a distinction between two types of factors that combine in the intensity. The first group of factors will encompass those that in some sense represent, or encode, information provided by the stimulus, such as objects, shapes, colors, edges, faces, etc.[2] We will denote the combination of these factors as $\lambda_2(x, y|s)$ and the integral

$$\Lambda_1 = \int\int \lambda_1(x, y|s)dxdy \tag{1}$$

will represent the total amount of information provided by the stimulus. The second group comprises factors that do not represent information of the stimulus but influence the attractiveness of the potential locations in another way, for example, heightening the intensity near the center of the stimulus would represent a central bias.

Combining Models for *When* and *Where*

The crucial part of the WALD-EM is how it relates the model for *when* and the model for *where* to each other. Recall that we conceptualize fixation duration as a period of evidence accumulation from a stimulus, and that information that provides this evidence is a part of the intensity maps. However, not all information is accessible at any single fixation (which is why we make saccades in the first place). Indeed, human vision is limited by the fact that only at the fovea, the place of the retina where the light falls from roughly around the center of gaze, great detail is available. This provides a key insight that the fixation durations should be dependent on how much information is available at the particular location the observer currently fixates. The physiological aspects of foveal, parafoveal, and extrafoveal vision are out of the scope of this article, but similar to other attempts for modeling of eye movements (Schütt et al., 2017; Schwetlick et al., 2019; Trukenbrod & Engbert, 2014), we represent the fact that vision is sharpest inside the fovea by implementing a so-called "attentional window." This window suppresses intensity of locations relatively farther from the center of gaze, and effectively limits the total information that is accessible to the observer given the current fixation location.

In essence, we define an attentional window as a function $a : \mathbb{R}^2 \rightarrow \mathbb{R} \in [0, 1]$ and denote it as $a(x, y|s)$, where s stands for the x and y coordinates of the current fixation. The value of a corresponds to the proportion of the intensity of locations at (x, y) given the current fixation location s. To get a representation of the actual intensity of different locations, given a particular fixation location s, we can multiply the intensity function by this attentional window:

$$\omega(x, y|s) = a(x, y|s) \times \lambda(x, y|s), \tag{2}$$

and the total amount of accessible intensity during a particular fixation s is $\Omega = \int\int \omega(x, y|s)dxdy$; the total amount of information accessible to the observer at a particular location will be denoted as

$$\Omega_1 = \int\int \omega_1(x, y|s)dxdy = \int\int a(x, y|s) \times \lambda_1(x, y|s)dxdy. \tag{3}$$

Figure 2 illustrates this concept with examples in one dimension.

The concept of attentional window is important in our model as it provides a link between the model for *when* and model for *where* to enable dependencies between the two. Specifically, we make the model of *when* depend on the model of *where*, and the attention window specifies how does that happen. In the following, we make this link explicit. In the model for *when*, the time it takes the observer to make a decision (to make a saccade) can be modeled as a Wald distribution with parameters drift ν and decision boundary α. However, it is likely that fixation durations vary depending on the surroundings of the current fixation location (Einhäuser et al., 2020; Nuthmann, 2017; Nuthmann et al., 2010).

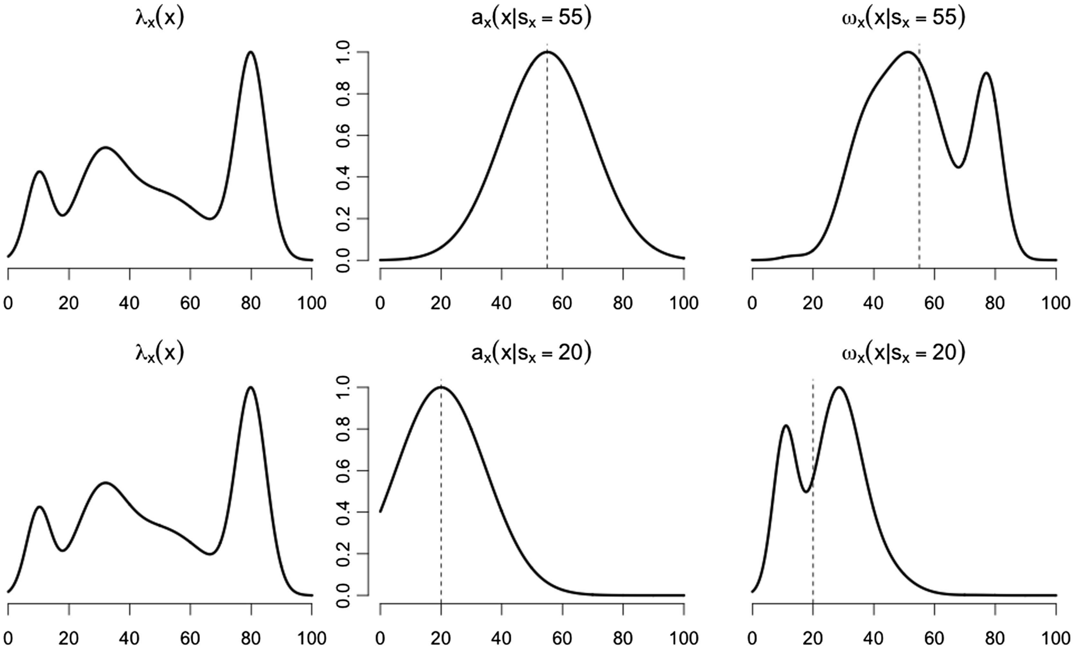
To link the model for *when* and *where*, we also make a distinction between factors that do and do not represent information provided by the stimulus, as we assume that only information has a potential to influence the fixation duration (e.g., fixating on a location particularly rich on detail will take longer on average than on a location with only a uniform background) and not other factors that do not provide information (e.g., central bias can attract people to make a saccade toward the center of the screen, but there is no immediately plausible mechanism for having longer fixation durations in the center of the screen compared to the edges just because it is in the center). Generally, the dependency of the fixation durations on fixation locations can be created in two ways, and the two approaches are discussed here.

In the first approach, we can assume that upon arriving at a location s, the observer harvests information from around that location with a drift rate ν, and once the information available from that location is depleted, the decision to *Go* is activated. In this framework, the total amount of information available through the attention window Ω_1 would replace the decision boundary α in the Wald model.

[2] We use the term *information* for a lack of a better word, and do not use it in a strict sense associated with the work of Shannon (1948).

Figure 2

The Left Panel Shows an Example Intensity Function λ(x) as a Function of Location Along the x-Coordinate. The Middle Panel Shows the Attention Window Given That the Current Fixation Is at $s_x = 55$ (Top) or $s_x = 20$ (Bottom). The Right Panel Shows the Intensity Accessible Through the Attention Window



In the second approach, we can adopt the idea from LATEST (Tatler et al., 2017) that the decision to make a saccade is based on continual comparison of two hypotheses (*Stay* vs. *Go*), where the "evidence" is based on the information provided if one or another decision is adopted. The evidence supporting the decision to stay is the total amount of information accessible through the attention window (Ω_1), whereas the evidence supporting the decision to Go is the total amount of information provided by the stimulus (Λ_1). In this framework, the drift rate of the Wald model equals the log of the ratio of the two evidences:

$$\nu = \ln\left(\frac{Go}{Stay}\right) = \ln\left(\frac{\Lambda_1}{\Omega_1}\right), \tag{4}$$

and the evidence accumulation continues until the decision threshold α is reached. The second approach is consistent with the increasing evidence that fixation durations are depending on a competition between the current and potential future fixation locations (Einhäuser et al., 2020). Crucially, both approaches share two main predictions: (a) increasing the width of the attention window increases the amount of information accessible through a single fixation, which has the effect of prolonging (on average) fixation durations, and (b) fixations in areas with a lot of information will have (on average) longer durations than fixations in areas with low information.

Concrete WALD-EM Model

In the previous section, we described the model in conceptual terms. However, in order to implement the model, there are several choices to be made about how to model the contribution of different factors, including their functional forms. Some of these choices will be purely pragmatic and statistical rather than theoretical, and are mostly motivated by the requirement to have a computationally tractable and modifiable model.

The model can be difficult to implement due to the two-dimensional integrals that are used to obtain the values of Λ_1 (total information on the stimulus) and Λ_1 (total information available through the attention window). The analytic tractability of these integrals relies on the functional form of the functions $\lambda(x, y)$ and $a(x, y)$, and consequently $\omega(x, y)$. This obstacle can be solved in two ways. However, these two approaches are not necessarily exclusive—later we apply a model combining both approaches. The two approaches we present here are not the only possible solutions, but are perhaps the most straightforward. Examples of other possible approaches are discussed in Gelman and Meng (1998), Wang and Wong (2007), and Azevedo-Filho and Shachter (1994).

First, it is possible to divide the stimulus into a grid of discrete locations, leading to an approximation of the continuous space, which leads to tractability regardless of the functional forms (i.e., integrals become sums) at the expense of loosing precision due to the discrete approximation. The degree of precision is arbitrary as it can be increased or decreased by changing the size of the cells in the grid, but could quickly lead to a computational bottleneck for fine-grained approximations due to the explosion of the number of terms to be summed.

Second, the construction of the functions at play can be carefully selected such that the integrals are analytically tractable. This avoids the need to specify the arbitrary precision of the discrete approach, and potentially leads to less computational burden. However, it may limit the flexibility of the model, as analytic solutions are possible only for a limited number of functional forms.

Modeling λ Lambda

The model for the function λ that converts the coordinates of the stimulus to intensity can be achieved in different ways. We generally desire to include different factors in the model, for example, central and directional biases, information about locations of objects on the scene, etc. This can be achieved by following (Barthelmé et al., 2013):

$$\lambda(x, y) = \Phi\left(\sum \beta_k f_k(x, y)\right), \tag{5}$$

where β_k is a weight of a factor k, f_k is a function that maps factor k to the locations (x, y), and Φ is analogous to a link function in GLMs. Particularly suitable candidates for this function are $\Phi(x) = \exp(x)$, $\Phi(x) = x$, $\Phi(x) = \ln(\exp(x) + 1)$ or their combinations (see Barthelmé et al., 2013, for the discussion of the differences between them).

In our application, we use Φ to be an identity function, which by using appropriate restrictions (specified below Equation 6) results in a mixture model:

$$\lambda(x, y) = \sum \pi_k f_k(x, y), \tag{6}$$

where $\pi_k \in [0,1]$ and $\sum \pi_k = 1$, $f_k(x, y) \geq 0 \forall x, \forall y$, and $\int\int f_k(x, y)dxdy = 1$, making the $\lambda(x, y)$ a proper probability density over a plane. The value of π_{k} then corresponds to the relative importance of a factor k, and $f_k(x, y)$ corresponds to a distribution of x and y locations under that factor. By definition, the value of $\Lambda = 1$ (total intensity of stimulus) for whatever setting of the parameters. A particularly attractive property of such definition is the fact that the separation between the factors that represent information on the stimulus from the factors that do not is straightforward. For example, if the first and second factors ($k = 1$ and $k = 2$) encode objects on the screen and saliency (which can plausibly play a role in fixation durations), whereas the third factor ($k = 3$) encodes a central bias (which arguably does not influence fixation durations) then we can simply drop the third factor from the calculations used in the model for fixation durations, and define $\lambda_1(x, y) = \pi_1 f_1(x,y) + \pi_2 f_2(x,y)$, and $\Lambda_1 = \pi_1 + \pi_2$.

Conceptually, a canonical interpretation of such formulation is that the mixture represents a generative model where the observer chooses the next fixation by first randomly selecting a factor k with probability π_k and then selects the location by randomly drawing from the density of the chosen factor f_k (Barthelmé et al., 2013). It is questionable whether this assumption is the most realistic—for example, taking $\Phi = \exp(x)$ (a log-additive model) corresponds to observers combining all factors into one meshed weighted map which determines the next fixation, an approach taken by Barthelmé et al. (2013)—the difference being that whereas the mixture model formulation allows to identify (with some probability) which particular factor was responsible for emitting a particular fixation, it is not the case for other models where all factors cause all fixations at the same time, but some have more influence than others. We believe that which approach is more realistic can be addressed by empirical comparison of different models that differ in these kinds of assumptions.

Calculating Ω

The crucial step is to determine the value of Ω (or Λ_1)—the total intensity available after filtering through the attention window $a(x,y|s)$. Recall that:

$$\Omega = \int\int \omega(x, y)dxdy = \int\int a(x, y)\lambda(x, y)dxdy. \tag{7}$$

Given the specification λ introduced in Equation 6, we can rewrite it as:

$$\begin{aligned}\Omega &= \int\int a(x, y)\sum \pi_k f_k(x, y)dxdy \\ &= \sum \pi_k \int\int a(x, y) f_k(x, y)dxdy,\end{aligned} \tag{8}$$

from which it is clearly visible that choice of the functional form of the attention window $a(x, y)$ and the individual factors $f_k(x, y)$ determine whether the model will be tractable without approximation through discretization. One of the possibilities to satisfy this is to model each $f_k(x, y)$ as a bivariate normal distribution, and $a(x, y)$ as a kernel of a bivariate normal distribution. Further, we will assume that the dimensions are uncorrelated, thus $f_k(x, y) = f_k(x) f_k(y)$ and $a(x, y) = a(x)a(y)$, where $f_k(.)$ is a Normal distribution with parameters μ_k and σ_k for the appropriate dimensions, and $a(.)$ is similarly the gaussian kernel with center at the current fixation (s) and scale parameter σ_a in the appropriate dimension. This allows us to rewrite the double integral in Equation 8 into a product of two integrals:

$$\Omega = \sum \pi_k \int a(x) f_k(x)dx \int a(y) f_k(y)dy, \tag{9}$$

which has a simple analytic solution:

$$\begin{aligned}
&\int a(x) f_k(x)dx \\
&= \int \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left[-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right] \exp\left[-\frac{(x-s_x)^2}{2\sigma_a^2}\right] dx \\
&= \frac{1}{\sqrt{2\pi\sigma_k^2}} \int \exp\left[-\frac{(x-\mu_k)^2}{2\sigma_k^2} - \frac{(x-s_x)^2}{2\sigma_a^2}\right] dx \\
&= \frac{1}{\sqrt{2\pi\sigma_k^2}} \int \exp\left[-\frac{\sigma_a^2(x-\mu_k)^2 + \sigma_k^2(x-s_x)^2}{2\sigma_a^2\sigma_k^2}\right] dx \\
&= \frac{1}{\sqrt{2\pi\sigma_k^2}} \int \exp\left[-\frac{(\sigma_a^2+\sigma_k^2)\left(x - \frac{\sigma_a^2\mu_k + \sigma_k^2 s_x}{\sigma_a^2+\sigma_k^2}\right)^2 + \frac{\sigma_a^2\sigma_k^2}{\sigma_a^2+\sigma_k^2}(\mu_k - s_x)^2}{2\sigma_a^2\sigma_k^2}\right] dx \\
&= \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left[-\frac{(\mu_k - s_x)^2}{2(\sigma_a^2+\sigma_k^2)}\right] \int \exp\left[-\frac{(\sigma_a^2+\sigma_k^2)\left(x - \frac{\sigma_a^2\mu_k + \sigma_k^2 s_x}{\sigma_a^2+\sigma_k^2}\right)^2}{2\sigma_a^2\sigma_k^2}\right] dx \\
&= \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left[-\frac{(\mu_k - s_x)^2}{2(\sigma_a^2+\sigma_k^2)}\right] \sqrt{2\pi} \sqrt{\frac{\sigma_a^2\sigma_k^2}{\sigma_a^2+\sigma_k^2}} \\
&= \frac{\sigma_a}{\sqrt{\sigma_a^2+\sigma_k^2}} \exp\left[-\frac{(\mu_k - s_x)^2}{2(\sigma_a^2+\sigma_k^2)}\right],
\end{aligned} \tag{10}$$

and equivalently in the y dimension.

Finally, we use the parametrization where the drift rate ν varies with the location of the fixation (Section 2.3). Combining the previous two equations, we can write the drift rate as follows:

$$\begin{aligned}\nu|s,\sigma_a,\lambda &= \ln(\Lambda_1) - \ln(\Omega_1)\\ &= \ln\sum_{k=1}^{K}\pi_k - \ln\sum_{k=1}^{K}\exp\left[\ln\pi_k + \sum_{i=1}^{2}\left(\ln\sigma_{ai} - \frac{\ln(\sigma_{ai}^2+\sigma_{ki}^2)}{2}\right.\right.\\ &\qquad \left.\left.- \frac{(\mu_{ki}-s_i)^2}{2(\sigma_{ai}^2+\sigma_{ki}^2)}\right)\right], \qquad (11)\end{aligned}$$

where the iteration over i only makes explicit the integration over x and y dimensions. The above expression was purposely written in the log-sum-exp-log form explicitly to bring it in line with its computational implementation (which is more stable in this form). In case not all factors provide information, the only change in Equation 11 would be the term K in the first summation (e.g., if the first two factors belong to λ_1 but not the third, $k = 3$ would be replaced with $k = 2$).

Likelihood

Assuming data in the form of $d \in \mathbb{R}_+^T$ as the durations and $s = (s_x, s_y) \in \mathbb{R}^{T\times 2}$ as the x and y coordinates of T observed fixations, the likelihood of the model can be written as:

$$\mathcal{L}(\theta|d,s) = \prod_{t=1}^{T}\lambda^{(t)}(s_x^{(t)}, s_y^{(t)}|\pi,\mu,\sigma) \times f_W(d^{(t)}|\nu^{(t)},\alpha), \qquad (12)$$

where the superscript for λ means that the intensity function might change during the course of time (which we show later), and $\nu^{(t}$ changes depending on the current location through Equation 11. f_W stands for the p.d.f. of the Wald distribution.

In general, assuming K factors included in the model, the model can have K parameters $\pi_1,\ldots,\pi_K$, $2 \times K$ parameters $\mu_1,\ldots,\mu_K$ (each a vector of 2 for x and y direction), $2 \times K$ parameters $\sigma_1,\ldots,\sigma_K$ (each a vector of 2 for the x and y directions), 2 parameters for σ_a (the width of the attention window in the x and y directions), and the decision boundary α, totaling $5K + 3 - 1$ free parameters. Depending on the actual factors included in the model, we will be able to fix or equate some parameters to reduce the number of parameters to be estimated, although it is not necessary to do so.

Including Saliency

An important branch of models that describe and predict distributions of fixation locations are saliency models. In our model, saliency can play a role as one of the factors determining the eye movement behavior. We define a saliency model (Itti et al., 1998; Itti & Koch, 2000) as an algorithm that takes the image (stimulus) as an input, and which produces an output, usually by assigning each pixel a value representing the local saliency of that pixel. Common features that these models consider important are local-global contrasts in color, intensity, and edges.

Saliency models enjoy a lot of success in predicting eye movement behavior and thus it seems reasonable to include some form of a saliency map as one of the factors in our model. Unfortunately, given the nature of the output of saliency models, it is not possible to implement the model fully analytically, and we will instead use discretization. To further reduce the computational complexity, we will reduce the resolution of the output of a saliency map.

Let's define a saliency map as **Sal**, where each of its element assigns a saliency to a pixel (in this context, pixels can be resized to contain multiple physical pixels of the display). Having I pixels in one dimension and J pixels in the other dimension, we have a total number of $P = I \times J$ pixels. We standardize the output of a saliency algorithm to ensure that $\sum_{p=1}^{P}\mathbf{sal}_p = 1$.

To include saliency into the model for *where*, we obtain a representation of the saliency on a continuous space of the x and y coordinates by defining the intensity function of saliency as a two-dimensional step function:

$$f(x,y) = \frac{\mathbf{sal}_{p(x,y)}}{h \times w}, \qquad (13)$$

where $p(x,y)$ returns the index of a pixel which is a super set of the position x and y, and where h and w is the height and width of the pixel. Standardization by the area of the pixel ensures that after converting the saliency map **Sal** to the intensity function, the volume $\int\int f(x,y)dxdy$ amounts to 1.

To include saliency into the model for *when*, we need to adopt additional simplifications as to evaluate the integral $\int\int a(x,y)\times f(x,y)dxdy$. We define x_p and y_p as the x and y coordinates of the center of a pixel p, respectively, and approximate $f(x, y)$ as:

$$f(x,y) \approx \sum_{p=1}^{P}\mathbf{sal}_p\text{Normal}(x|x_p,\kappa)\text{Normal}(y|y_p,\kappa), \qquad (14)$$

which leads to (using results in Equation 8):

$$\begin{aligned}&\int\int a(x,y)f(x,y)dxdy\\ &\approx \sum_{p=1}^{P}\mathbf{sal}_p\frac{\sigma_a^2}{\sigma_a^2+\kappa}\exp\left(-\frac{(x_p-s_x)^2+(y_p-s_y)^2}{2(\sigma_a^2+\kappa)}\right), \qquad (15)\end{aligned}$$

which we can further simplify by letting $\kappa \to 0$:

$$\int\int a(x,y)f(x,y) \approx \sum_{p=1}^{P}\mathbf{sal}_p\exp\left(-\frac{(x_p-s_x)^2+(y_p-s_y)^2}{2\sigma_a^2}\right). \qquad (16)$$

These steps enable us to approximately compute the drift rate by substituting the discrete saliency map with a continuous function.

However, this implementation still requires serious computational resources: for example, fitting a model that includes a saliency map of resolution of 800 × 600 pixels would mean summing up $P = 800 \times 600 = 480{,}000$ terms for every fixation in every iteration of the fitting procedure.

There are generally three ways to alleviate the problem of computational complexity. First, it is possible not to include the discrete factor in the model for *when*, but only include it into the factor for *where*. However, leaving it out does not solve the problem, but rather avoids it altogether. Second, it is possible to downsample the output of the saliency map. Indeed, many saliency algorithms already output the saliency map that has a resolution smaller than the original image (e.g., by a factor of 16 in each of the dimensions, Itti et al., 1998). Having an input image of dimensions of 800 × 600 pixels then leads to quite substantial reduction: Instead of summing up 480,000 terms we need to sum up only about 2,000. Downsampling the saliency maps to have smaller resolution than the input image is also desirable from a measurement perspective as the eye-tracking devices likely have measurement error that translates to several pixels of the input image. Downsampled saliency maps then correspond better to the level of precision of the data. Third, it is possible to limit the summation only for the terms that lie in a relative proximity from the current fixation. For example, the attention window lets through only at most 1.1% of the total weights of the pixels that lie at a distance of $3\sigma_a$ or more, essentially meaning that many of the terms in the sum are basically zero. Leaving out the pixels that are that far from the current fixation can reduce the number of terms to be summed by a great amount while not sacrificing much of the computational accuracy. Downsampling and limiting the summations are not mutually exclusive and can be used at the same time—an approach we take in the practical implementation of our model.

Application: Infant Scene Viewing

Here, we apply a particular realization of the model to data by van Renswoude et al. (2020) to demonstrate its use in applied context. This data set was originally collected with the aim to investigate the role of bottom-up versus top-down factors in infants' eye movements. Specifically, the data set was collected to assess whether object familiarity is associated with specific patterns in eye movement behavior of infants when looking at pictures of real-world scenes. In the following application, we are interested in the extent to which four different factors influence the distribution of fixation locations and the timing of saccades.

The four factors that we considered are the (a) locations (and sizes) of objects on the scene (van Renswoude et al., 2020; Xu et al., 2014), (b) saliency (Itti & Koch, 2000, 2001; Itti et al., 1998), (c) exploitation (i.e., tendency to make repeated fixations in a relative proximity to a previous fixation; Malem-Shinitski et al., 2020), and (d) central bias (van Renswoude et al., 2019).

The model was fitted on half of the data set (with the other half used for a following cross-validation). To accommodate individual differences between participants, we generalized the model using hierarchical modeling (the details are explained below). As such, we obtain assessment of the individual differences between the participants in terms of their tendency to dwell longer on current locations (captured by the decision boundary), and the width of their attention window.

Data Descriptive

The data contains recordings of 47 participants looking at 29 static pictures selected from the pool of 700 images created by Xu et al. (2014). 39 participants looked at all 29 stimuli (min = 5, mean = 27.6, median = 29, max = 29 viewed images per participant). The mean number of fixations per trial was 11.4 (*SD* = 4.3); the total number of fixations in the data set is 14,807.

We split the data set into two parts, one of which we used to estimate the parameters of a model (number of fixations = 7,207), and one of which we used to validate the predictions of the model. We counter balanced the number of trials per participant in the two sets to ensure that both data sets contain some data from all participants and all items. The details of the procedure are available at https://github.com/Kucharssim/WALD-EM/blob/master/scripts/prepare_data.md.

The exact form of splitting the data in this way had the following reasons. First, the aim of this article is not to generalize findings, but rather as a conceptual proof of concept—that the model is applicable to eye-tracking data and captures some interesting patterns in the data. Second, being able to generalize to a population is contingent on additional requirements besides cross-validation procedure, such that the participants and stimuli were randomly selected from the target population. We did not define which population of infants we would like to generalize to, and do not assess whether they indeed represent that population. Further, we *know* for certain that the 29 stimuli are not randomly selected from the pool of 700 images by Xu et al. (2014), making it difficult to generalize even to this pool, and even more problematic to some more general population of static images. Lastly, in terms of our goals, our procedure is more robust against differences between the train and test sets caused by randomly splitting small sets of data. Usually, for smaller data sets, procedures such as k-fold cross-validation (or leave-one-out) is usually done for this purpose, compared to split half cross-validation. However, *k*-fold cross-validation was not an option due to the computational demands of the model. Thus, by giving up aspirations for generalizing our findings, our splitting procedure ensured that we could still perform cross-validation, but ensuring that potential problems of cross-validation are not caused by randomly choosing an "outlying" participant or stimulus into the train or test sets.

Initial Model

The model contains four factors that determine the fixation locations, two of which are included in the model for fixation duration.

The model for *where* is composed of four factors, and so we can describe the distribution of fixation locations as follows:

$$(x, y) \sim \sum_{k=1}^{4} \pi_k f_k(x, y|\theta_k), \tag{17}$$

where π_k are the weights of different factors and f_k is the distribution of a factor k with parameters θ_k.

The first factor is the location and sizes of objects on the scene. We assume that each object on the scene can have different level of attractivity and that larger objects distribute their total attractivity over larger area. This idea can be expressed by another mixture:

$$f_1(x, y|\theta_1) = \sum_j \omega_j \text{Normal}(x|\text{center}_{xj}, \gamma \times \text{width}_j)\text{Normal}(y|\text{center}_{yj}, \gamma \times \text{height}_j), \tag{18}$$

where ω_j are the individual attractivities of different objects on a particular image. Parameter γ is a scaling factor that stretches or compresses the attractivity of objects proportionally to their sizes.

The second factor is the saliency, which we treated as described in Equation 13.

The third factor can be described as an exploitation factor, and captures the phenomenon that people tend to linger close to the current fixation location: we model it as a bi variate normal distribution centered at the fixation location at time t to predict the fixation location at time $t + 1$:

$$f_3(x, y|\theta_3) = \text{Normal}\ (x|s_x^t, \sigma_e)\ \text{Normal}\ (y|s_y^t, \sigma_e) \tag{19}$$

The fourth factor represents the central bias, and is modeled as a bi variate normal distribution centered at the center of the screen ($x_c = 400,\ y_c = 300$)

$$f_4(x, y|\theta_4) = \text{Normal}(x|400, \sigma_d)\text{Normal}(y|300, \sigma_d) \tag{20}$$

For the fixation durations, we only consider the first two factors influential: the latter two factors do not stand for *information* presented on the screen, but rather spatial biases, and therefore should not have any influence on saccade timing.

The model for fixation duration can be summarised as follows:

$$d \sim \text{Wald}(\nu, \alpha)$$

$$\nu = \ln(\pi_1 + \pi_2) - \ln\left(\iint a(x, y|\sigma_a)\left(\pi_1 f_1(x, y|\theta_1) + \pi_2 f_2(x, y|\theta_2)\right)dxdy\right).$$

We also modeled the individual differences between participants by modeling their decision boundary and width of the attention window as random terms. Because both of these parameters need to be positive, we modeled them on the log scale:

$$\ln(\alpha_i) = \mu_\alpha + z_i * \sigma_\alpha$$

$$z_i \sim \text{Normal}(0,1),$$

where α_i stands for the decision boundary of participant i, and μ_α with σ_α are the estimated group mean and standard deviation of the parameter α on the log scale. The same approach was taken for the attention window σ_a.

We used weakly informative priors on the parameters that were based on prior predictive simulations done when building the model (Schad et al., 2019). The priors are accesible at the model file: https://git.io/JfjuJ.

We implemented the model using the probabilistic programming language Stan (Carpenter et al., 2017) interfacing with R (R Core Team, 2020) using the package rstan (Guo et al., 2020; Stan Development Team, 2020). The following additional R packages were used to produce the output (in no particular order): OpenImageR (Mouselimis, 2019), Rcpp (Eddelbuettel & Balamuta, 2017; Eddelbuettel et al., 2020), ggforce (Pedersen, 2019a), gtools (Warnes et al., 2020), here (Müller, 2017), imager (Barthelme, 2020), knitr (Xie, 2015, 2020), patchwork (Pedersen, 2019b), plotrix (Lemon et al., 2019), pracma (Borchers, 2019), tidybayes (Kay, 2020), and tidyverse (Wickham, 2019; Wickham et al., 2019). Throughout the development of the model, we conducted simulation studies to validate our implementation. The results are documented in the following folder of our project repository https://github.com/Kucharssim/WALD-EM/tree/master/documents. The current (small) simulation results are encouraging in terms of parameter recovery, but we did not invest our resources into a full validation study due to the computational demands of the model.

Results—Initial Model

We ran 10 MCMC chains with random starting values and default tuning parameters set by Stan. Each chain ran for 1,000 warm up and 1,000 sampling iterations, resulting in a total of 10,000 samples used for inference. The model ran without any divergent transitions. We examined the potential scale reduction factor $\hat{R}$, trace plots, auto correlations, and the number of effective samples to identify potential problems with convergence. We did not find indications of poor convergence, and thus proceed with interpreting the model.

Posterior Predictive Checks

We generated posterior predictives for the data set used for estimating the parameters, and for the hold out data set, to assess whether the fitted model reproduces the observed patterns in the data, and whether the patterns that the model picks up from the data carry over to the hold out data set. This enables us to contrast features that are desirable to be captured by the model (i.e., patterns that are present both in the fitting and hold out data set), from features that are not so desirable to be captured by the model (i.e., patterns that are present in fitting but not hold out data set).

The model is able to capture the characteristic distribution of the fixation durations, as documented in Figure 3, although the model predicts a slightly fatter right tail than that of the data. We also inspected the model's predictions of the fixation durations for individual participants, to assess whether it captures their individual differences. Figure 4 shows that the model is well able to capture individual differences between participants in respect to their fixation durations.

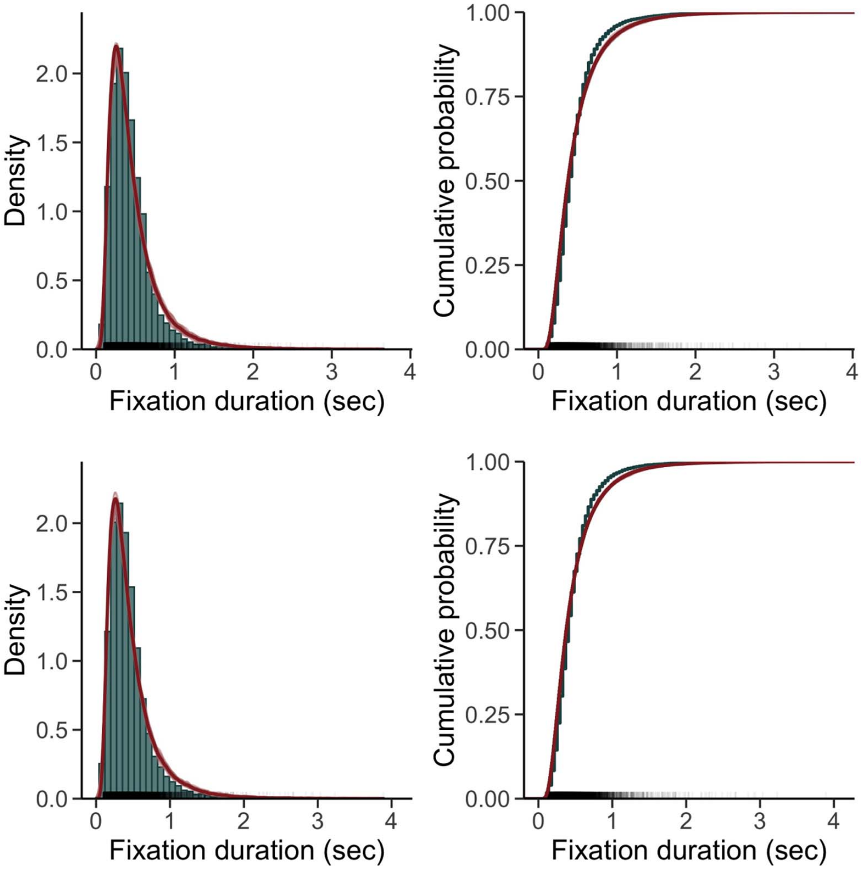
The model also reproduces the distributions of fixation locations. Figure 5 shows an example for one particular stimulus (image number 251 from Xu et al., 2014). The top-right in Figure 5 displays the four factors included in the model, which combine proportionally to their weights to the posterior predictive distribution (labeled as predicted fixations).

Next to the variables used to fit the data (fixation durations and fixation locations), we also checked on other quantities implied by the model. Specifically, we checked whether we can reproduce the distribution of saccade amplitudes and the distribution of saccade angles (as these derivative measures provide additional insights into the model performance, for example, Henderson & Hayes, 2018). Saccade amplitude was measured as the Euclidean distance between two successive fixations in units of pixels. Saccade angle was calculated as an angle in radians between the horizontal axis of the screen and the vector that connects two successive fixations.

Figure 6 shows the observed versus predicted distributions of saccade amplitudes on one example stimulus (shown in Figure 5). The model usually captures the distribution of saccade amplitudes relatively well, exhibiting two modes. Figure 6 shows the observed versus predicted distributions of saccade amplitudes on the same stimulus. The prediction of saccade angles is relatively good, as the model picks up patterns of saccade directions specific to a specific stimulus. As an illustration, Figure 7 shows that the model captures saccade directions in the top-right and bottom-left directions on the

Figure 3

Predicted (Red) Versus Observed (Green) Distribution of the Fixation Durations; Predictions Using the Initial Model. Left Panel Shows Histogram of the Empirical Data Versus the Density Estimate Using Gaussian Kernel of the Posterior Predictives. Right Panel Shows Empirical Cumulative Distribution Functions. Top Panel Shows the Data Used for Fitting the Model, Bottom Panel the Hold out Data Used for Cross-Validation



Note. See the online article for the color version of this figure.

stimulus shown in Figure 5. However, overall the model does not capture well excess of saccades in the horizontal direction (see Figure 12), which could be an indication that the model needs to be expanded with a factor that represents a horizontal bias (van Renswoude et al., 2016).

Parameter Estimates

The results indicated that the most important factor driving fixations was the locations of objects on the scene (weight = 0.37, 95% CI [0.35, 0.40]), followed by exploitation (weight = 0.30, 95% CI [0.28, 0.31]), saliency (weight = 0.18, 95% CI [0.16, 0.20]), and central bias (weight = 0.16, 95% CI [0.14, 0.17]).

The parameter that controls sizes of objects as identified by Xu et al. (2014) indicated that people fixate relatively close to the centroids of the objects (scale = 0.23, 95% CI [0.22, 0.24]). The exploitation region had a standard deviation $\sigma = 34.58$ (95% CI [33.15, 36.06]) pixels, whereas the central bias region had a standard deviation $\sigma = 93.84$ (95% CI [88.65, 98.91])) pixels. Relatively speaking, central bias is less focused than the exploitation factor.

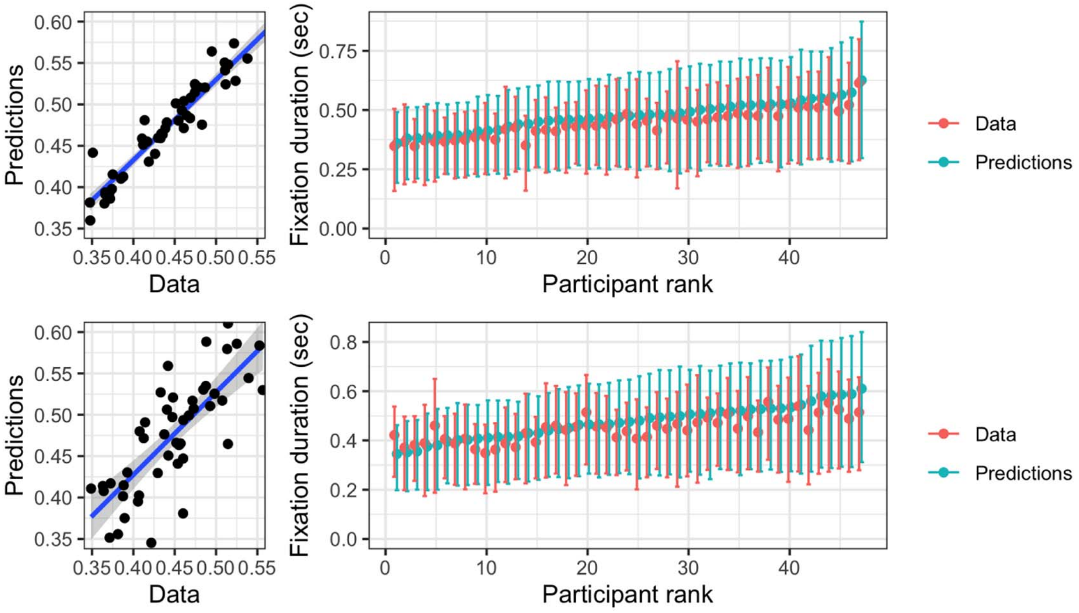
Extended Model

The original model fared well capturing the distribution of fixation durations and the overall distribution of fixation locations, and was able to a small degree to capture an excess of horizontal saccades without this being explicitly built into the model. However, the discrepancy between the model's predictions and data show that the tendency to make horizontal saccades is particularly noteworthy and possibly needs an extra explanation.

To explore whether we can improve the model's capability to reproduce the amount of saccades in the horizontal direction, we extended the model. Specifically, we added another factor into the model for fixation locations, representing the horizontal bias. To create a factor that represents a saccadic bias (such as horizontal bias), instead of location preferences, it is possible to transform fixation locations (x and y coordinates) into a saccade representation (angle θ and amplitude r of a saccade):

Figure 4

Posterior Predictive Checks for the Individual Differences in Fixation Durations. Left Panel Shows the Observed (-Axis)x and Predicted (-Axis)y Mean Fixation Durations for Each Participant Separately. Right Panel Shows the Observed (Red) Versus Predicted (Blue) Mean Fixation Durations, 20% and 80% Quantiles (Whiskers). The Top Panel Shows the Results in Training Set, Bottom Panel Shows the Results in the Test Set



Note. See the online article for the color version of this figure.

$$\begin{aligned} \theta &= \arctan\left(\frac{\Delta y}{\Delta x}\right) \\ r &= \sqrt{\Delta x^2 + \Delta y^2}, \end{aligned} \tag{21}$$

where $\Delta x = x^t - x^{t-1}$ and $\Delta y = y^t - y^{t-1}$ represent a fixation as the difference of the x and y coordinates compared to the previous fixations (we set $x^0 = 400$ and $y^0 = 300$ as that is the middle of the screen).

That way, we can substitute a factor of locations with a factor of saccade angles and amplitudes:

$$f(x, y) = \frac{f(\theta, r)}{r}, \tag{22}$$

where the denominator r is the Jacobian determinant representing the stretching of the space after the change of variables from cartesian to polar coordinates: $dxdy = rdrd\theta$.

To create the joint density of the angle and amplitude, we express it using the chain rule of probability:

$$f(\theta, r) = f(\theta) \times f(r|\theta). \tag{23}$$

The important part of this factor is the distribution of saccade angles, for which we specify the following distribution:

$$f(\theta) = 0.5 \text{ vonMises } (0, \kappa) + 0.5 \text{ vonMises } (\pi, \kappa), \tag{24}$$

which specifies a mixture of von Mises distributions with centers fixed to 0 and π (i.e., right and left direction, respectively), and a concentration κ which is estimated from the data. The mixture weights are fixed to 0.5 as we assume that saccades in the left direction are equally attractive as saccades to the right direction.

The conditional density $f(r|\theta)$ is chosen to be a uniform stretched over the interval between 0 and the maximum saccade length that would not fall outside of the screen if it was launched from the position (x^{t-1}, y^{t-1}) under the direction θ.

The generative mechanism for such a joint density is the following. First, the observer draws a saccade angle from the distribution $f(\theta)$. Then, the observer draws a point along a line under the sampled angle θ, that goes between location (x^{t-1}, y^{t-1}) and the edge of the screen. This point is the new fixation position.

Figure 8 shows and example of the function $f(\theta, r)$ on the screen coordinates, with (x^{t-1}, y^{t-1}) set to the center to the screen, and $\kappa = 15$.

The rest of the model stayed the same.

Results—Extended Model

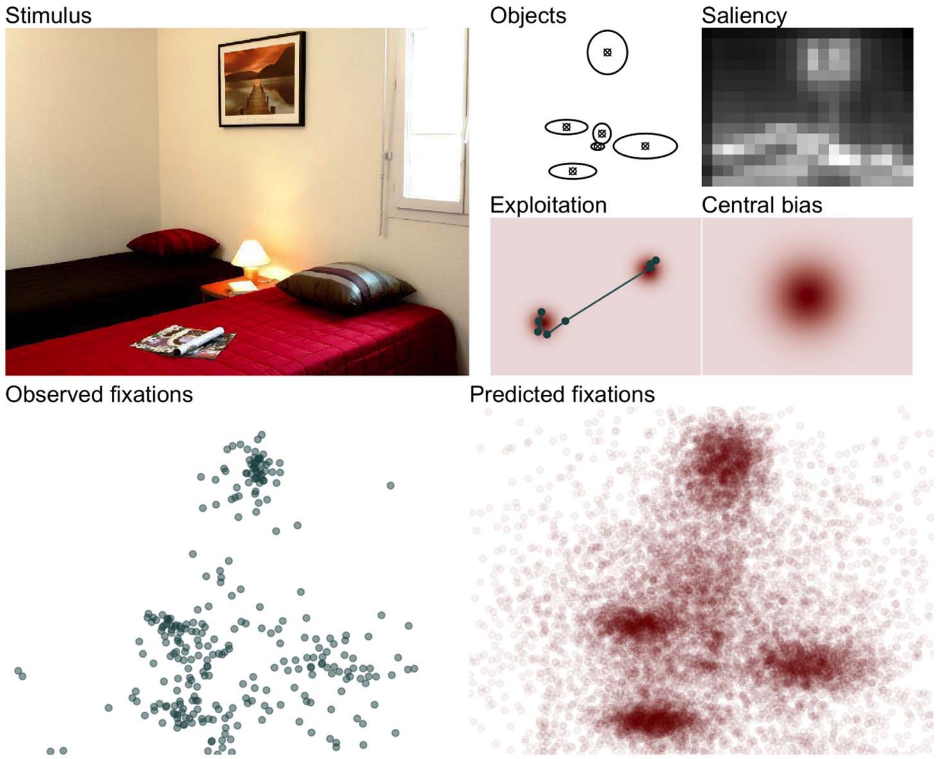
We fitted the extended model in the same way as the initial model: We ran 10 MCMC chains with random starting values and default tuning parameters set by Stan. Each chain run for 1,000 warm up and 1,000 sampling iterations, resulting in a total of 10,000 samples used for inference. The model ran without any divergent transitions. We examined the convergence diagnostics, to find that we could not identify potential problems with convergence. Thus, we proceed with interpreting the model.

Posterior Predictive Checks

We conducted posterior predictive checks the same way as with the previous model: Comparing the predicted and observed distribution of fixation durations, fixation locations, saccade amplitudes, and saccade angles. The extended model performed similar to

Figure 5

Example of the Model for Fixation Locations on the Stimulus Created by Xu et al. (2014; Top Left Panel). The Four Factors Influencing the Fixation Locations are Depicted in the Top Right Quadrant. The Bottom Left Panel Shows the Observed Fixation Locations, and the Bottom Right the Draws From the Posterior Predictive Distribution of the Fitted Model



Note. See the online article for the color version of this figure.

the initial model in terms of the first three variables (see Figures 9 and 11). As Figure 12 demonstrates, the extended model did better in terms of reproducing the overall distribution of saccade angles—being able to reproduce the excess of saccades going in the horizontal direction better after we have explicitly added a factor that represents horizontal bias. However, it is still visible that there is still some potential to improve the model predictions. Figure 10 shows that the model is well able to capture individual differences between participants in respect to their fixation durations.

Model Comparison Using Cross-Validation

To assess whether the extended model did better at predicting the data compared to the initial model, we computed the log-likelihood of the hold-out set under the two models, given the posterior distributions of the parameters. This way, we obtained distributions of the log-likelihood for the two models based on their out-of-sample performance which we use for cross-validating the models. To compare the two distributions, we computed the distribution of the log-likelihood differences: $\Delta \log \mathcal{L} = \log \mathcal{L}(\text{Model2}) - \log \mathcal{L}(\text{Model1})$: Positive values mean that the extended model predicted the hold-out data better than the initial model, and negative values mean that the initial model predicted the hold-out data better than the initial model.

The log-likelihood difference distribution (median = 45.77, IQR [15.18, 77.06]) indicated that the extended model was better at predicting the hold-out data than the intitial model: adding horizontal bias into the model increased the model's predictive success.

Parameter Estimates

The estimates indicated that the most important factor were the objects on the scene (weight = 0.35, 95% [0.33, 0.37]), followed by exploitation (weight = 0.31, 95% CI [0.29, .32]), saliency (weight = 0.14, 95% CI [0.13, 0.16]), central bias (weight = 0.13, 95% CI [0.11, 0.15]), and lastly the horizontal bias (weight = 0.07, 95% CI [0.06, 0.8]); Tables 1 and 2).

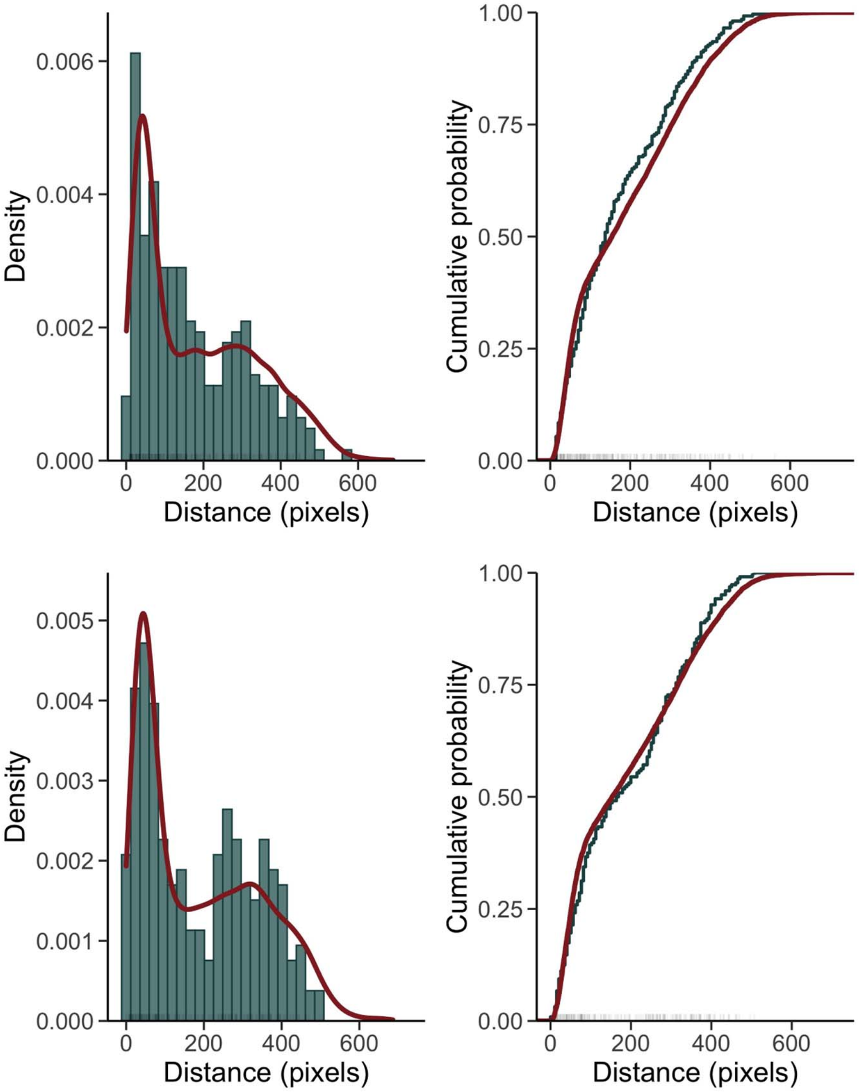
The parameters that control the individual factors were very similar to those of the initial model. The parameter that controls sizes of objects indicated that people fixate relatively close to the centroids of the objects (scale = 0.23, 95% CI [0.22, 0.24]). The exploitation region had a scale (σ = 34.28, 95% CI [32.80, 35.78]) of about a third of that of the central bias (σ = 98.68, 95% CI [93.42, 103.99]). The additional parameter that controls the concentration of the horizontal bias was estimated to κ = 18.36, 95% CI [13.4, 24.11].

Benefits of Joint Modeling of Fixation Locations and Fixation Durations

The application of the model presented in the article showed that the model is able to fit a particular data set relatively well.

Figure 6

Observed (Blue) Versus Predicted (Red) Saccade Amplitude on One Particular Stimulus; Predictions Using the Initial Model. Top Panel Shows the Data Used for Fitting the Model, Bottom Panel the Hold out Data Used for Cross-Validation



Note. See the online article for the color version of this figure.

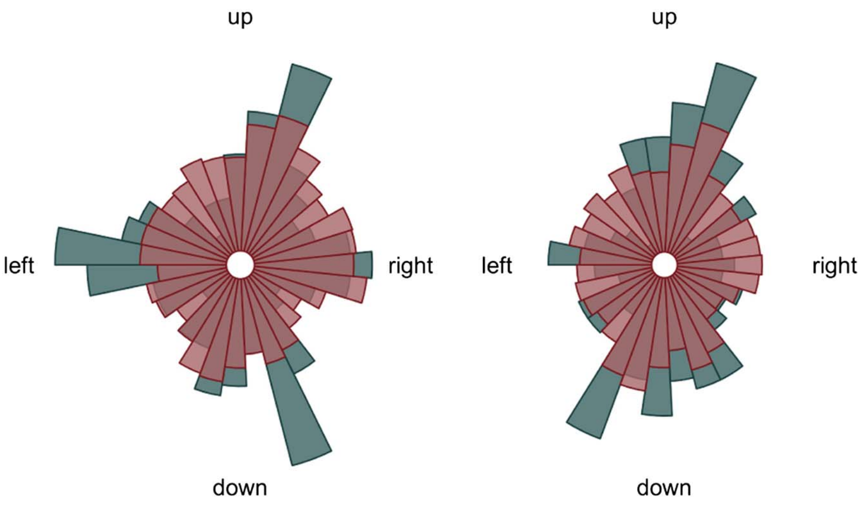
The strength of our approach is that it can model both fixation durations and fixation locations concomitantly. This has two general benefits compared to models that consider fixation durations and fixation locations separately. In this section, we discuss these benefits more explicitly.

First, having different sources of data gives us more information to estimate parameters of interest. For example, one could fit some variant of a mixture model to fixation locations to estimate the importance of various factors that influence eye movements. However, in our model, the weights of these factors not only play a role for the fixation locations, but also come into play when calculating the drift rate of the Wald process, therefore they are informed by the fixation durations as well. This benefits parsimony of our models as well as introduces the potential to estimate parameters with greater precision, allowing us to even estimate parameters that would have been otherwise hardly identifiable.

Second, modeling fixation locations and durations together enables us to capture some dependencies between the two. Potentially, this could lead to modeling phenomena that occur in both spatial and temporal dimensions. In the case of the current approach, the model has a built-in global dependency between locations and durations due to the way it represents their joint probability. Specifically, the distribution of fixation durations depends both on the individual characteristics of the observer (by having two parameters vary between participants), but also on the surroundings of the current fixation. The surroundings of the observers' fixation are taken into account when evaluating the drift rate of the Wald process, where the intensity function is passed through the attention window.

Figure 7

Observed (Blue) Versus Predicted (Red) Saccade Angle on One Particular Stimulus. Plot on the Left Shows the Data Used for Fitting the Model, Plot on the Right Shows the Hold out Data Used for Cross-Validation



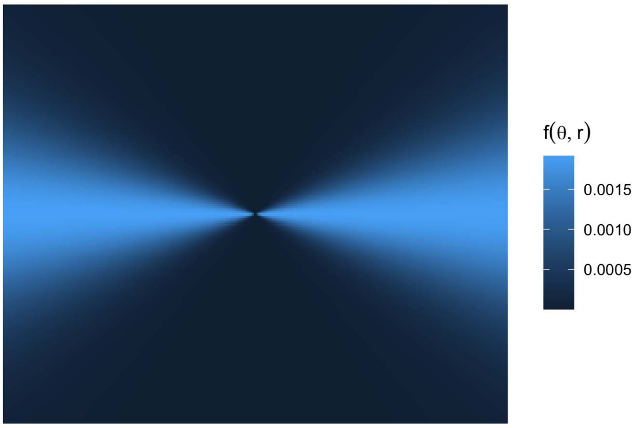
Note. See the online article for the color version of this figure.

The intensity function is concurrently the function that (stochastically) determines fixation locations. Thus, the model makes the prediction that fixations on locations that are more likely to be fixated than others will (on average) be longer than locations that are relatively less likely to be fixated. Conceptually, locations with a lot of information will have a lot of attraction, and lead to more fixations and longer durations.

To test that the model's predictions are accurate in this respect, we derive the following two hypotheses[3] that follow from the model's predictions. (a) The model is able to predict individual fixation durations. Because the model already explains individual differences in fixation durations, we will look into the correlation between predicted and observed fixation duration for each participant separately. A positive correlation within a participant suggests that the model is able to pick up information around the fixation location to adjust its prediction about the fixation duration. (b) Both in the data and in the model, there exists a positive correlation between how often (or how likely) a particular area on the image is fixated, and a mean fixation duration in that area. To test this hypothesis, we split images into a grid and calculate the correlations for each image separately (there might be differences between images that we left unmodelled). For both hypotheses, we used the Bayesian inference scheme for correlations developed by Ly et al. (2018). We present the results both for the training set, the test set, and with the two sets combined. Here we show the results from the extended model. The results from the initial model are nearly identical.

Figure 8

Example of the Joint Density of Saccade Angle and Saccade Amplitude Plotted on the Screen Dimensions. The Density Highlights Saccades in the Left and Right Directions Relative to the Current Fixation (In This Figure, the Center of the Screen), Representing the Horizontal Bias



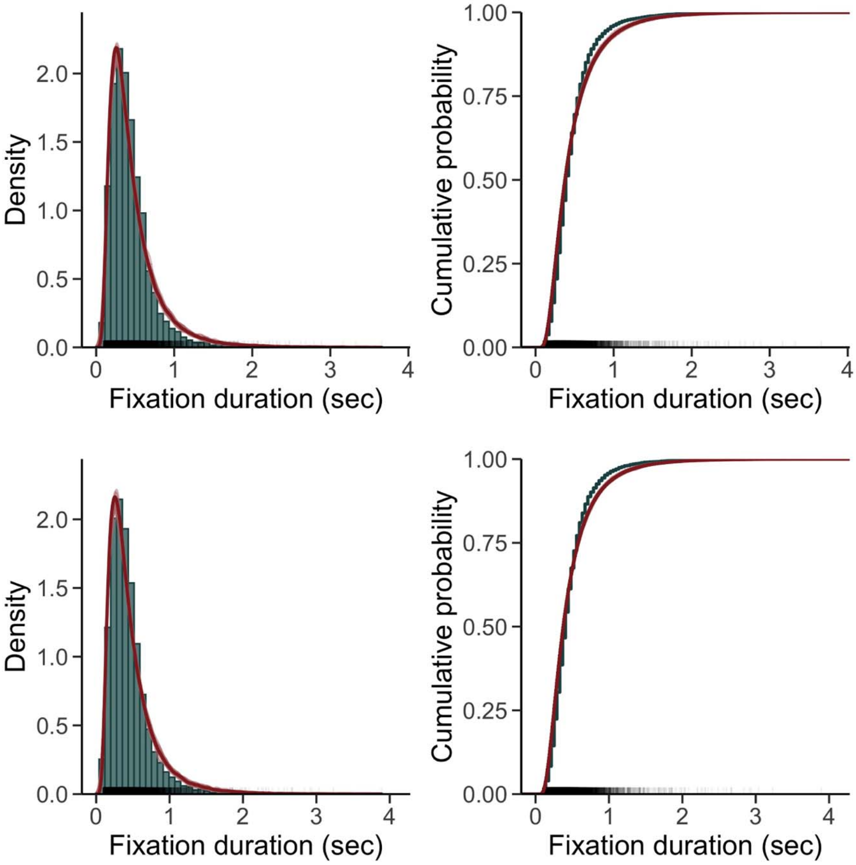
Note. See the online article for the color version of this figure.

Figures 13 and 14 show the results related to hypothesis (a) that the model's predictions positively correlate with the observed fixation durations. Figure 13 shows the scatterplot between the mean of the predictive distribution for each fixation duration and the observed fixation durations, superimposed by the regression line for each participant separately. Figure 14 shows the corresponding Pearson's correlation coefficients and their Bayes factors (testing null hypothesis of no correlation versus alternative hypothesis of positive correlation). We calculated the same for the train set ("In sample"), the test set ("Out of sample"), and the two data set combined ("Combined"). Overall, most of the correlations were positive, suggesting that the model is able to pick up some information about the surroundings of a fixation to inform the fixation duration. However, the correlations were relatively low and some correlations remained in the region where the Bayes factor does not

[3] N.B. these hypotheses were formulated post hoc, based on the feedback from an anonymous reviewer to whom we are thankful for this idea.

Figure 9

Predicted (Red) Versus Observed (Green) Distribution of the Fixation Durations; Predictions Using the Extended Model. Left Panel Shows Histogram of the Empirical Data Versus the Density Estimate Using Gaussian Kernel of the Posterior Predictives. Right Panel Shows Empirical Cumulative Distribution Functions. Top Panel Shows the Data Used for Fitting the Model, Bottom Panel the Hold out Data Used for Cross-Validation



Note. See the online article for the color version of this figure.

strongly prefer either the null or the alternative, suggesting that much of the variability of fixation durations is yet to be explained by additional mechanisms.

Figures 15 and 16 show the results related to hypothesis (b) that there is a relationship between fixation duration and fixation probability, that is, that locations that are more frequently visited are also fixated with longer durations. To assess this hypothesis we first split each image into a grid of 50 × 50 pixels, leading to 16 × 12 cells in the grid in each image. For each cell, we calculated the probability of fixating inside of it using the model's predictive distribution for the fixation locations, and the predicted and observed mean fixation durations. As before, we conducted this calculation for the train set, test set, and combined data. Figure 15 shows the scatterplot of the log probability of fixating a particular cell of the grid and the log mean fixation duration, with superimposed regression line for each image separately. The top panel shows the relation that was found in the data, the bottom panel shows the relation that is reproduced by the model. Figure 16 further shows the observed Pearson's correlation and the corresponding Bayes factors testing the null hypothesis of no correlation versus the alternative hypothesis of positive correlation. For the majority of images (except for three images that show correlations near zero), there appears to be a positive relation between probability of fixating a particular location and the mean fixation duration at that location. Arguably the relationship is stronger in the model than in the data (see Figure 15). However, one needs to keep in mind that the mean observed fixation durations are noisy because they are often calculated from only a couple of fixations inside a particular cell of the grid. Whether this is a sufficient explanation or there is additional model misspecification

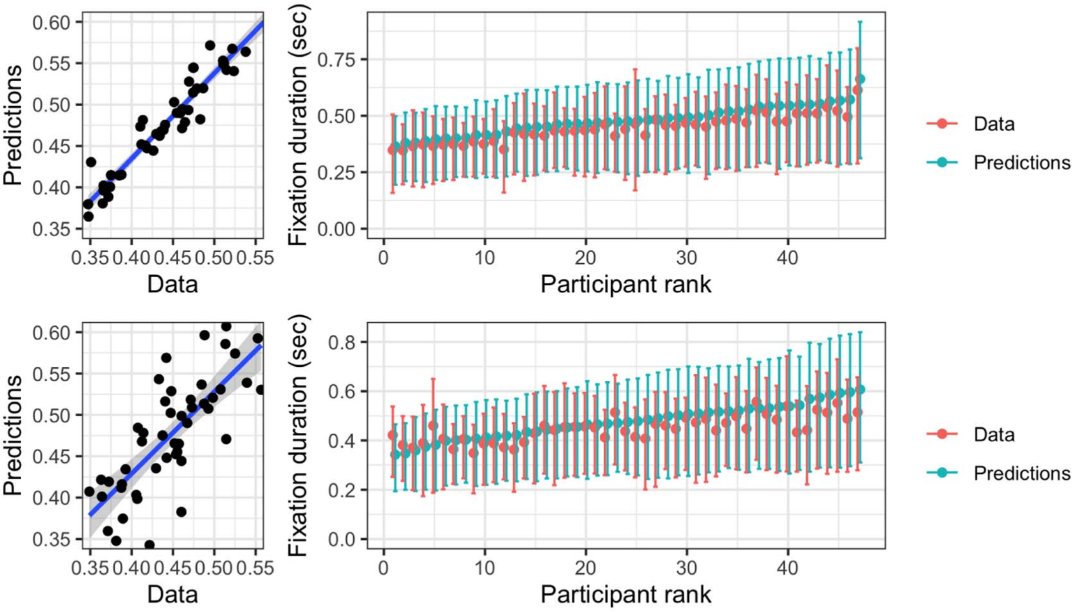
Table 1

Descriptives of the Posterior Distribution of the Factor Weights Under the Initial Model

			Quantiles	
Factor	Mean	*SD*	2.5%	97.5%
Objects	0.36	0.01	0.34	0.39
Saliency	0.17	0.01	0.16	0.19
Exploitation	0.33	0.01	0.32	0.35
Central bias	0.13	0.01	0.11	0.15

Figure 10

Posterior Predictive Checks for the Individual Differences in Fixation Durations. Left Panel Shows the Observed (-Axis)x and Predicted (-Axis)y Mean Fixation Durations for Each Participant Separately. Right Panel Shows the Observed (Red) Versus Predicted (Blue) Mean Fixation Durations, 20% and 80% Quantiles (Whiskers). The Top Panel Shows the Results in Training Set, Bottom Panel Shows the Results in the Test Set



Note. See the online article for the color version of this figure.

that causes this discrepancy is a potentially interesting avenue for future research. Taken together, these results demonstrate that the model is able to capture some global dependency of fixation durations on the attractivity of location of the image.

Conclusion and Discussion

This article presents arguments that theoretically grounded statistical models are important for validating predictions of the emerging theoretical framework against observed phenomena as well as detecting new empirical phenomena to be explained by said theory. Our model is grounded in the theoretical assumptions that can be verbally summarized as follows: (a) fixation durations depend on observers harvesting information from the stimulus, which is a noisy accumulation process, (b) saccades are launched when the observer concludes that staying at the current location is no longer advantageous over moving to another location, (c) picking a new location depends on an internal "intensity" map over the stimulus, which is a combination of different "factors," such as information on the screen or for example systematic tendencies that highlight certain locations in contrast to others, and (d) observers harvest information from the relative proximity of the center of gaze, subjected to the limitations of their visual acuity—an assumption that provides the link between fixation durations and fixation locations. We consider these the core theoretical assumptions of the model. From this listing of assumptions, it should be evident that we are relatively more vague on the mechanism behind selection of the location of the next saccade. This is because the model offers flexibility by making use of "factors" that influence this selection, and because these "factors" on their own can represent different theoretical viewpoints. For example, original saliency models, such as that of Itti et al. (1998), can be considered a theory of fixation selection of itself, as it describes the rise of saliency map as neurons firing according to surround-background differences in image intensity, contrast of colors, and orientation of edges. We think this is a strength of our model as it allows to "plug-in" different explanations of the data without having to heavily modify the model.

Table 2

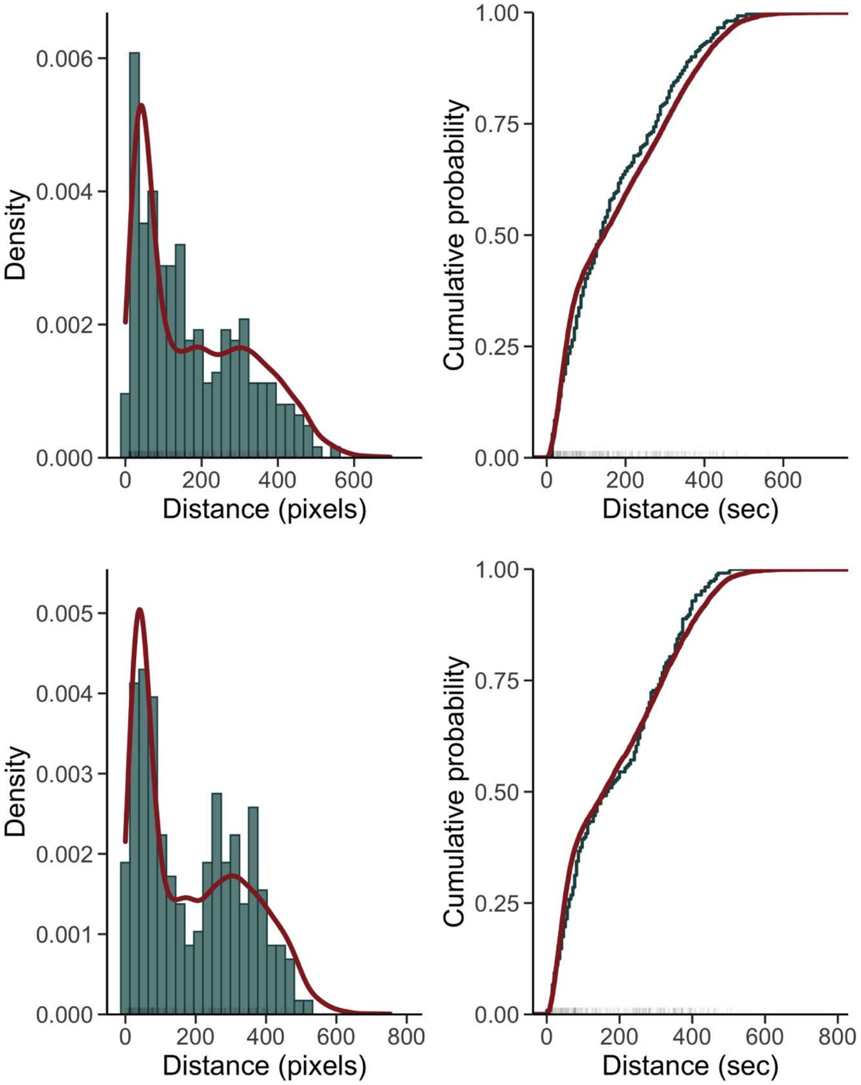
Descriptives of the Posterior Distribution of the Factor Weights Under the Extended Model

			Quantiles	
Factor	Mean	*SD*	2.5%	97.5%
Objects	0.35	0.01	0.33	0.37
Saliency	0.14	0.01	0.13	0.16
Exploitation	0.31	0.01	0.29	0.32
Central bias	0.13	0.01	0.11	0.15
Horizontal bias	0.07	0.01	0.06	0.08

In this article, we developed a model to analyze eye movement data by specifying a joint probability distribution of the fixation duration and fixation locations. To our knowledge, this is the first attempt to model fixation durations and fixation locations by defining a joint likelihood function of these two random variables. Using Bayesian inference, we were able to fit and extend the model such that the predicted patterns of the fixation durations and fixation locations align very closely with those of the observed data. Drawing upon the strengths of specifying models using likelihood functions (Schütt et al., 2017), we demonstrated how to diagnose,

Figure 11

Observed (Blue) Versus Predicted (Red) Saccade Amplitude on One Particular Stimulus; Predictions Using the Extended Model. Top Panel Shows the Data Used for Fitting the Model, Bottom Panel the Hold out Data Used for Cross-Validation



Note. See the online article for the color version of this figure.

improve, and compare models so that they capture phenomena of interest present in real data. An example application showed that adding horizontal bias to the model improved the model's ability to capture the distribution of saccade angles.

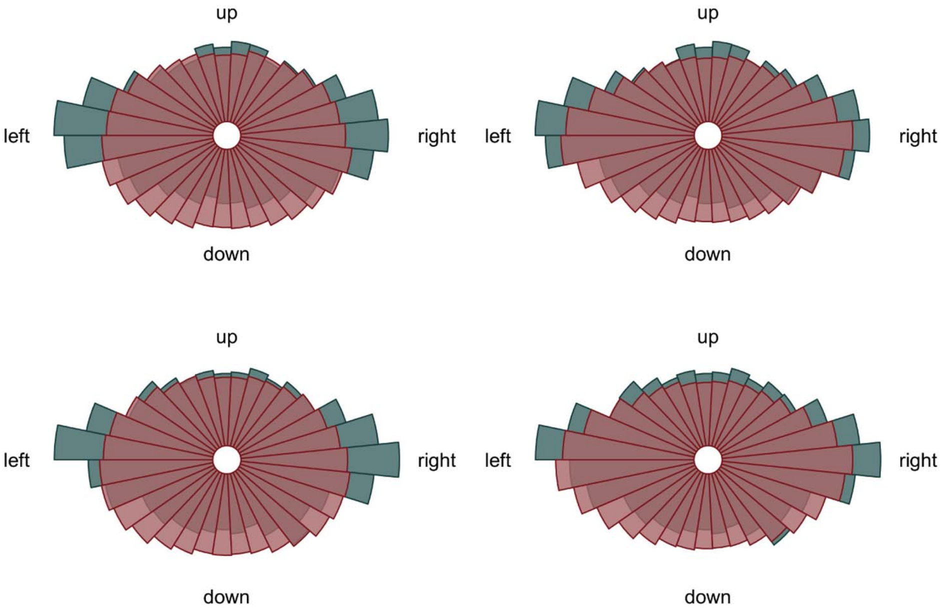
The advantage of this approach is that it is possible for the model to be fitted to data (given that it is a statistical model with a likelihood function) and used to generate new data that can be contrasted with observed phenomena (such as distribution of saccade angles). In case the model does not perform well in capturing these phenomena, it can be iteratively modified or improved until the model does so, or is ultimately rejected. Further, the model provides a relatively straightforward interpretation of the model parameters, facilitating the inference and possibly theorizing about the underlying mechanisms.

In our application, the results suggested that the most important factors determining eye movement behavior are the locations of objects on a scene, immediately followed by the tendency to make repeated fixations in a location nearby the current fixation. Salience and central bias had lower importance, and horizontal bias the least, although all factors made a significant contribution to fitting the data.

With respect to the central and horizontal biases, there is an ongoing debate on what is exactly the cause of these phenomena. Possible explanations range from being caused entirely by the image content (e.g., objects mostly aligned in the center of the images, or objects mostly aligned along horizontal axes or the horizon), to being some sort of interaction between image content and systematic bias toward centers or horizontal saccades, to being completely explained by systematic tendencies, caused by physiological,

Figure 12

Predicted Versus Observed Distribution of Saccade Angles Under the Initial (Left) and Extended (Right) Model, Over All Stimuli in the Data Set; The Top Panel Displays the Data Set Used to Fit the Model, Bottom Displays the Hold out Data Set



Note. See the online article for the color version of this figure.

learned or strategic aspects (Foulsham et al., 2013, 2018; Le Meur & Liu, 2015; Tatler & Vincent, 2008; Tseng et al., 2009; van Renswoude et al., 2016, 2019). It is also possible that these three sources of the observed "biases" are not mutually exclusive. The model would be able to generate some central and horizontal bias with only the objects and saliency factors, representing the first category. In our model, we ended up using additional central bias and horizontal bias factors that were modeled as completely independent of the image content, hence representing the third category (independent of image content). Including these additional factors improved the fit of the model above factors that encode the image content, lending some credit to the third type of explanations. However, apart from the need of replicating this finding on other data sets, one needs to also implement central and horizontal biases that explicitly interact with the image content. Then, it will be possible to test all these explanations against each other.

In this article, our focus was mainly on free scene viewing, and so was our example. We hope and believe that the current model can be adapted to different contexts as well, as is it so easily modified that it can include different factors or possibly various terms that accommodate various experimental designs or research questions. For example, it should be possible to use the model to compare demographic (e.g., adults versus infants) or experimental groups (e.g., free viewing instructions versus visual search instructions), providing alternatives to already established analytic methods (e.g., Coutrot et al., 2018), or even adopt the model to specific purposes—such as strategic influences on eye movements in cognitive tasks (e.g., Kucharský et al., 2020) or economic games (e.g., Polonio et al., 2015).

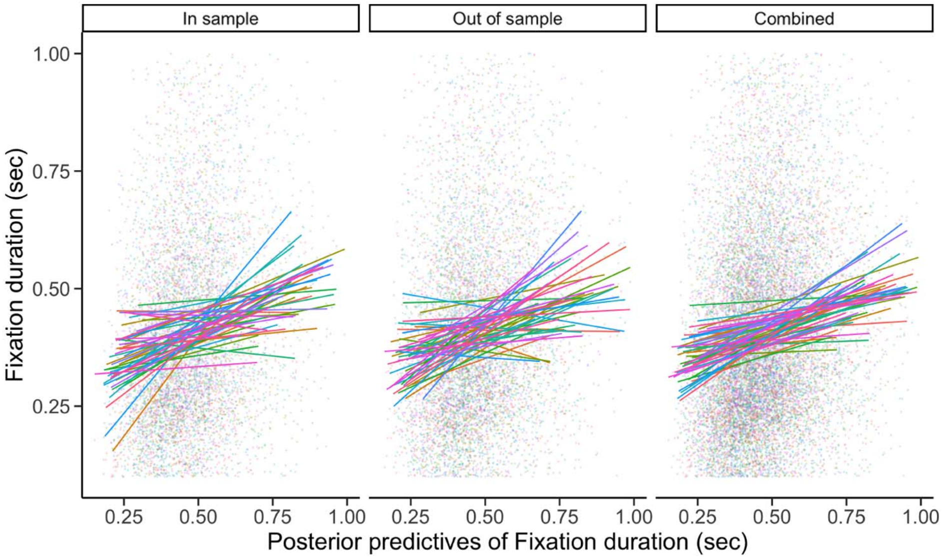
This article followed an unusual strategy in model comparison, and that leads to some considerations regarding generalizability of our findings. The strategy of splitting the data set into two parts allows us to assess the adequacy of the models to describe patterns in the data that were not used for fitting the model. Thus, cross-validation procedures (e.g., model comparison) are possible. However, the splitting procedure ensured that for each trial, whether or not included in the train or test sets, the model has some information about the participant (from other trials done by that participant) and the stimulus (from other participants on that stimulus) in that trial. Thus, the data in the test set cannot be considered completely *out of sample* in the traditional sense, which is one of the requirements for generalizability (next to additional assumptions). The cross-validation still gives us information about overfitting, but does not aim for generalizing to a new population. This means that if we talk about one model fitting better than another model, we mean it is better at capturing patterns in the current sample of participants looking at the current sample of stimuli.

Extensions and Future Directions

Although the final model fits relatively well, there are plenty of ways to make it better in the future. For example, previous research suggested that the central bias is slightly more stretched in the horizontal compared to the vertical dimension (Clarke & Tatler, 2014; Tatler, 2007). In our application, we hold the width of the central bias in two dimensions equal. This could have created a slight misfit of the central bias factor, and could also underestimate the model's ability to produce saccade angles in the horizontal direction. Further, we hold the widths of factors in the model constant wherever possible to make the simplest model we could apply do the data, and so this limitation can relate also to the exploitation factor and the attention window (both of which we

Figure 13

The Mean of the Posterior Predictive for the Fixation Durations (x-Axis) vs the Observed Fixation Durations (y-Axis) for Each Participant Separately. The Scatterplot Suppresses Outliers Above 1 s. Regression Lines Are Superimposed to Highlight the Variability Between Participants



Note. See the online article for the color version of this figure.

assume is spherical). We also modeled all factors as independent normal distributions. In general, this assumption is not strictly necessary, and could be relaxed by specifying the components as bivariate normal, that is, estimating their correlations. Luckily, these issues can be solved easily in case the data indicate to do so.

Potential model misspecification could also arise from modeling the horizontal bias. It has been shown that von Mises distribution is not necessarily optimal for describing the distribution of the saccade angles, due to the fact that the real distributions of saccade angles are typically more peaked than what von Mises distribution allows

Figure 14

The Correlations Between Predicted Mean Fixation Duration and Observed Fixation Duration for Each Participant Separately and the Log Bayes Factors Testing the Null Model (No Correlation) vs The Alternative (positive Correlation, Specified by Stretched β (10, 10) Truncated at Zero). The Dotted Lines Show the Region of "Anectdotal" Evidence (Bayes Factor Between 1/3 and 3), That is, There is Not Enough Information to Say Anything Meaningful About Presence or Absence of the Correlation

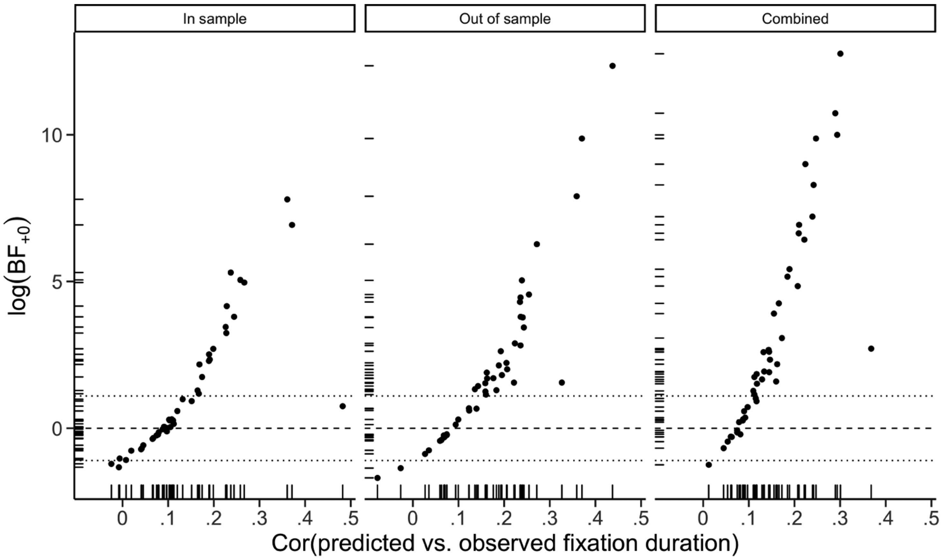
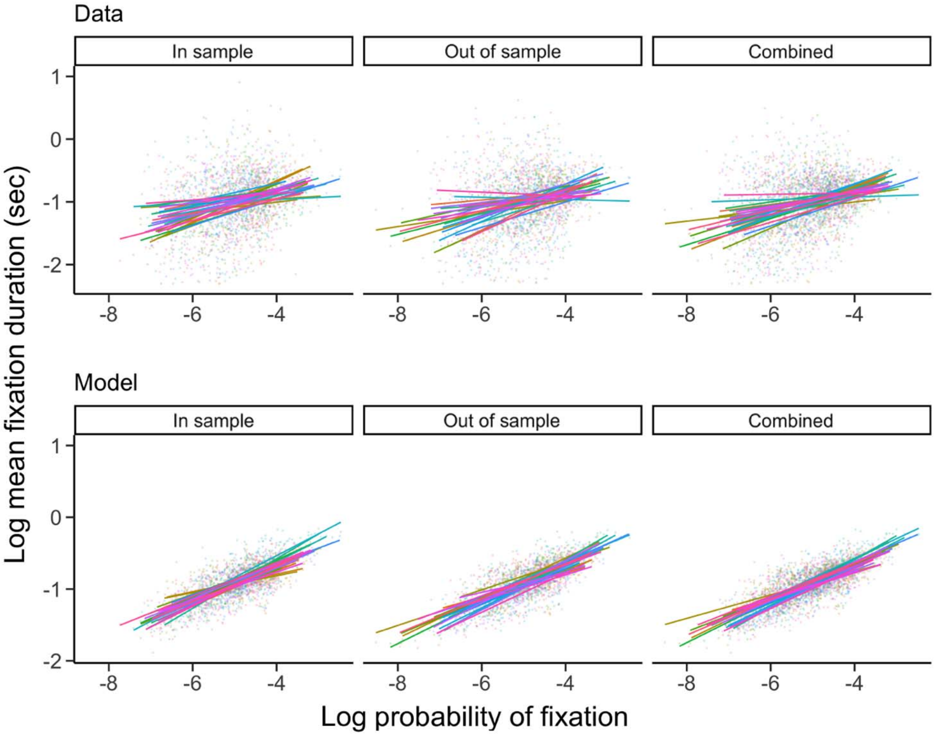


Figure 15

The Correlations Between the Log of the Predicted Probability of Fixation and the Log of the Mean Fixation Duration of Cells in the Grid. Superimposed are the Regression Lines Per Each Image Separately. Regression Lines are Superimposed to Highlight the Variability Between Items. The Top Panel Shows the Relationship in the Data, the Bottom Panel Shows the Relationship Reproduced by the Model



Note. See the online article for the color version of this figure.

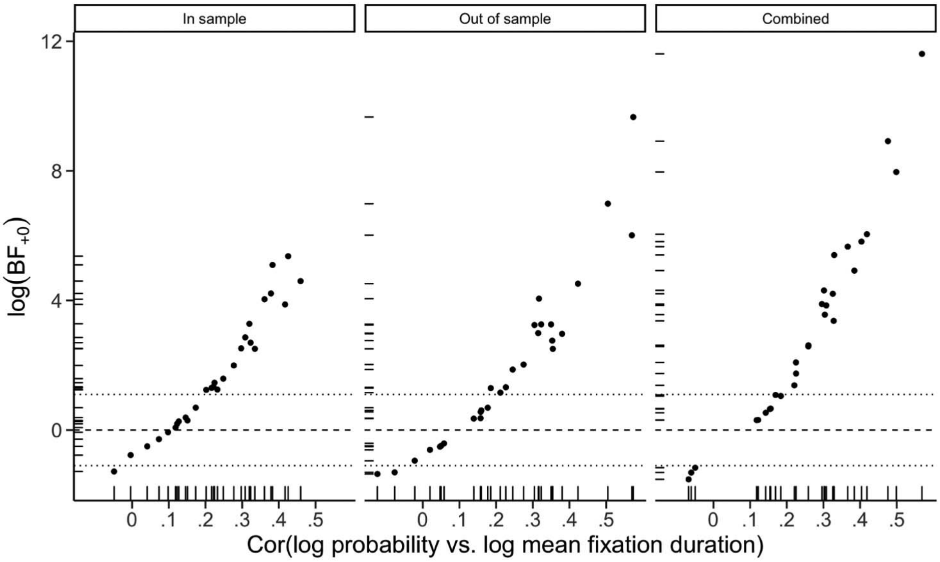
(Mulder et al., 2020). We used the von Mises distribution because it is relatively well known and can be fitted easily in Stan, whereas alternative distributions—such as the power Batchelet distribution as proposed by Mulder et al. (2020)—would make the implementation much more complicated. A second potential misspecification of the horizontal bias could be that the current implementation assumes that at any point in time, it is equally likely to make a saccade to the left direction as to right direction. However, this is likely not true, as intuitively we could think that having a fixation very close to a left border of the scene would lead to a rightward saccade with a very high probability (Clarke et al., 2017). This assumption could have underestimated the weight of the horizontal bias contribution compared to the other factors.

Additional model misspecifications could arise from modeling many parameters as fixed across participants and stimuli. In the current model, we only modeled the most obvious source of individual differences—the width of the attention window and the decision boundary—as random between participants. Importantly, these parameters only affect fixation durations, therefore the current model cannot capture individual differences in selecting fixation locations. However, it is probable that to better account for the patterns in the data (and to justify generalizibility to a population of observers and a population of stimuli; Yarkoni, 2019), we would need to model many of the currently fixed parameters as random. For example, it is desirable to assume that participants can differ in the weights of the different factors, or that the importance of different factors are different even in different stimuli. Being able to generalize beyond the current data set was however not the focus of this article. However, the model is relatively flexible and including parameters as random should be possible in future applications. However, as explained above, the aim of this article is not a generalization per se, for this reason, the current model is perhaps not unreasonably simplistic. Additionally, allowing the model to capture random effects does not automatically ensure generalizability, and additional effort and checks need to be put in place when designing the experiment.

The current implementation of the model does not capture differences between images. This means that one is not able to investigate the effect of the differences of total saliency or other factors, such as the number of objects on the screen, between different images on the eye movement behavior. This is due to the constraint that (a) the factor weights sum to 1 and are set equal between images and (b) observed factors (such as saliency) are normalized before entering the model. It is possible to either relax these constraints or add more parameters accounting for the differences between images, which is a candidate for future extensions. However, careful development and validation of different approaches should precede the application. As such, it is not presented in this article.

Figure 16

The Correlations Between the Log of the Predicted Probability of Fixation and the Log of the Mean Fixation Duration of Cells in the Grid, for Each Image Separately on the x-Axis and the Corresponding Log Bayes Factors Testing the Null Model (no Correlation) vs the Alternative (Positive Correlation, Specified by Stretched β (10, 10) Truncated at Zero). The Dotted Lines Shows the Region of "Anectdotal" Evidence (Bayes Factor Between 1/3 and 3), That is, There is Not Enough Information to Say Anything Meaningful About Presence or Absence of the Correlation



It is possible that the proposed mechanism underlying the model's architecture will need to be adapted in the future. For example, our assumption is that observers linger on a current fixation for the time it takes to decide to move to another location. It is possible that a different mechanism drives fixation durations. We also assume that the time it takes to select a new fixation location, and plan and execute the saccade is zero, that observers plan new target fixations only one step ahead (ignoring preplanned saccades), or that once a decision to make a saccade is made, there is no stopping in launching it—assumptions that were relaxed in different modeling approaches (Nuthmann, 2017; Nuthmann et al., 2010; Trukenbrod & Engbert, 2014). We also assume that different factors combine in an additive manner (and can only increase the intensity), which may not be a realistic assumption (Barthelmé et al., 2013)—for example, a typical factor that is plausibly affecting fixation locations is the inhibition of return, which inhibits intensity of locations that were already visited (Klein, 2000). We believe that such alternative conceptual ideas could be contrasted with the current model by developing new mathematical and statistical models that concretely implement these. Having specific models that are derived from concrete theoretical assumptions will hopefully facilitate our understanding of the real generative mechanisms (Borsboom et al., 2020; Schütt et al., 2017) that are relevant in eye-movement research.

We believe that similar attempts to modeling eye movements can influence both experimental practice as well as the theoretical advancements in the eye-tracking research. We made our code available online (https://github.com/Kucharssim/WALD-EM), along with additional materials that provide details about building and applying the model, so that other researchers can seek inspiration and help, if they wish to use our ideas for furthering their own work. Additional work should be done on the front of model validation through more extensive simulation studies. We hope that the current model will eventually be superseded by a better one—which would be a good sign of a healthy progress of our scientific understanding of visual perception. In the meantime, we hope that the proposed model will spark interest in applied and theoretical research of eye movements and provide valuable insights.

References

Anders, R., Alario, F. X., & van Maanen, L. (2016). The shifted Wald distribution for response time data analysis. *Psychological Methods*, *21*(3), 309–327.

Azevedo-Filho, A., & Shachter, R. D. (1994). Laplace's method approximations for probabilistic inference in belief networks with continuous variables. In R. L. Mantaras & D. Poole (Eds.), *Uncertainty proceedings 1994* (pp. 28–36). Elsevier. https://doi.org/10.1016/B978-1-55860-332-5.50009-2

Barthelme, S. (2020). *imager: Image processing library based on 'cimg'* (R package version 0.42.1) [Computer software manual]. https://CRAN.R-project.org/package=imager

Barthelmé, S., Trukenbrod, H., Engbert, R., & Wichmann, F. (2013). Modeling fixation locations using spatial point processes. *Journal of Vision*, *13*(12), Article 1. https://doi.org/10.1167/13.12.1

Borchers, H. W. (2019). *pracma: Practical numerical math functions* (R package version 2.2.9) [Computer software manual]. https://CRAN.R-project.org/package=pracma

Borsboom, D., van der Maas, H., Dalege, J., Kievit, R., & Haig, B. (2020). Theory construction methodology: A practical framework for theory formation in psychology. *PsyArXiv*. https://doi.org/10.31234/osf.io/w5tp8

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive* *psychology*, *57*(3), 153–178. https://doi.org/10.1016/j.cogpsych.2007.12.002

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*(1), 1–32. https://doi.org/10.18637/jss.v076.i01

Carpenter, R. H., & Williams, M. (1995). Neural computation of log likelihood in control of saccadic eye movements. *Nature*, *377*(6544), 59–62. https://doi.org/10.1038/377059a0

Chen, S.-C., She, H.-C., Chuang, M.-H., Wu, J.-Y., Tsai, J.-L., & Jung, T.-P. (2014). Eye movements predict students' computer-based assessment performance of physics concepts in different presentation modalities. *Computers & Education*, *74*, 61–72. https://doi.org/10.1016/j.compedu.2013.12.012

Chhikara, R., & Folks, L. J. (1988). *The inverse Gaussian distribution: Theory, methodology, and applications*. CRC Press.

Clarke, A. D. F., Stainer, M. J., Tatler, B. W., & Hunt, A. R. (2017). The saccadic flow baseline: Accounting for image-independent biases in fixation behavior. *Journal of Vision*, *17*(11), Article 12. https://doi.org/10.1167/17.11.12

Clarke, A. D. F., & Tatler, B. W. (2014). Deriving an appropriate baseline for describing _xation behaviour. *Vision Research*, *102*, 41–51. https://doi.org/10.1016/j.visres.2014.06.016

Coutrot, A., Hsiao, J. H., & Chan, A. B. (2018). Scanpath modeling and classi_cation with hidden Markov models. *Behavior Research Methods*, *50*(1), 362–379. https://doi.org/10.3758/s13428-017-0876-8

De Haas, B., Iakovidis, A. L., Schwarzkopf, D. S., & Gegenfurtner, K. R. (2019). Individual differences in visual salience vary along semantic dimensions. *Proceedings of the National Academy of Sciences*, *116*(24), 11687–11692. https://doi.org/10.1073/pnas.1820553116

Duchowski, A. T. (2007). *Eye tracking methodology: Theory and practice*. Springer.

Eddelbuettel, D., & Balamuta, J. J. (2017, August). Extending extitR with extitC++: A Brief Introduction to extitRcpp. *PeerJ Preprints*, *5*, Article e3188v1. https://doi.org/10.7287/peerj.preprints.3188v1

Eddelbuettel, D., Francois, R., Allaire, J., Ushey, K., Kou, Q., Russell, N., Bates, D., & Chambers, J. (2020). *Rcpp: Seamlessr and c++ integration* [Computer software manual]. http://www.rcpp.org, http://dirk.eddelbuettel.com/code/rcpp.html, https://github.com/RcppCore/Rcpp

Einhäuser, W., Atzert, C., & Nuthmann, A. (2020). Fixation durations in natural scene viewing are guided by peripheral scene content. *Journal of Vision*, *20*(4), Article 15. https://doi.org/10.1167/jov.20.4.15

Findlay, J. M., & Walker, R. (1999). A model of saccade generation based on parallel processing and competitive inhibition. *Behavioral and Brain Sciences*, *22*(4), 661–674. https://doi.org/10.1017/s0140525x99002150

Forstmann, B. U., & Wagenmakers, E.-J. (2015). *An introduction to model-based cognitive neuroscience*. Springer.

Foulsham, T., Frost, E., & Sage, L. (2018). Stable individual differences predict eye movements to the left, but not handedness or line bisection. *Vision Research*, *144*, 38–46. https://doi.org/10.1016/j.visres.2018.02.002

Foulsham, T., Gray, A., Nasiopoulos, E., & Kingstone, A. (2013). Leftward biases in picture scanning and line bisection: A gaze-contingent window study. *Vision Research*, *78*, 14–25. https://doi.org/10.1016/j.visres.2012.12.001

Gelman, A., & Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, *13*(2), 163–185.

Guo, J., Gabry, J., & Goodrich, B. (2020). *rstan: R interface to stan* (R package version 2.19.3) [Computer software manual]. https://CRAN.R-project.org/package=rstan

Hayes, T. R., Petrov, A. A., & Sederberg, P. B. (2015). Do we really become smarter when our fluid-intelligence test scores improve? *Intelligence*, *48*, 1–14. https://doi.org/10.1016/j.intell.2014.10.005

Henderson, J. M., & Hayes, T. R. (2018). Meaning guides attention in realworld scene images: Evidence from eye movements and meaning maps. *Journal of Vision*, *18* (6), Article 10. https://doi.org/10.1167/18.6.10

Henderson, J. M., Nuthmann, A., & Luke, S. G. (2013). Eye movement control during scene viewing: Immediate effects of scene luminance on fixation durations. *Journal of Experimental Psychology: Human Perception and Performance*, *39*(2), 318–322. https://doi.org/10.1037/a0031224

Itti, L., & Borji, A. (2014). Computational models: Bottom-up and top-down aspects. In A. C. Nobre & S. Kastner (Eds.), *Oxford handbook of attention*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199675111.013.026

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40* (10–12), 1489–1506. https://doi.org/10.1016/S0042-6989(99)00163-7

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203. https://doi.org/10.1038/35058500

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259. https://doi.org/10.1109/34.730558

Kay, M. (2020). *tidybayes: Tidy data and 'geoms' for bayesian models* (R package version 2.0.3) [Computer software manual]. https://CRAN.R-project.org/package=tidybayes

Klein, R. M. (2000). Inhibition of return. *Trends in cognitive sciences*, *4*(4), 138–147.

Kucharský, Š., Visser, I., Tru escu, G.-O., Laurence, P. G., Zaharieva, M., & Raijmakers, M. E. J. (2020). Cognitive strategies revealed by clustering eye movement transitions. *Journal of Eye Movement Research*, *13*(1), 1–20. https://doi.org/10.16910/jemr.13.1.1

Le Meur, O., & Liu, Z. (2015). Saccadic model of eye movements for freeviewing condition. *Vision research*, *116*, 152–164. https://doi.org/10.1016/j.visres.2014.12.026

Lemon, J., Bolker, B., Oom, S., Klein, E., Rowlingson, B., Wickham, H., Tyagi, A., Eterradossi, O., Grothendieck, G., Toews, M., Kane, J., Turner, R., Witthoft, C., Stander, J., Petzoldt, T., Duursma, R., Biancotto, E., Levy, O., Dutang, C., & Venables, B. (2019). *plotrix: Various plotting functions* (R package version 3.7-7) [Computer software manual]. https://CRAN.R-project.org/package=plotrix

Ly, A., Marsman, M., & Wagenmakers, E.-J. (2018). Analytic posteriors for Pearson's correlation coefficient. *Statistica Neerlandica*, *72*(1), 4–13.

Malem-Shinitski, N., Opper, M., Reich, S., Schwetlick, L., Seelig, S. A., & Engbert, R. (2020). A mathematical model of exploration and exploitation in natural scene viewing. *BioarXiv*. https://doi.org/10.1101/2020.04.16.044677

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan* (2nd ed.). CRC press.

Mouselimis, L. (2019). *Openimager: An image processing toolkit* (R package version 1.1.6) [Computer software manual]. https://CRAN.R-project.org/package=OpenImageR

Mulder, K., Klugkist, I., van Renswoude, D. R., & Visser, I. (2020). Mixtures of peaked power batschelet distributions for circular data with application to saccade directions. *Journal of Mathematical Psychology*, *95*(102309), 1–12. https://doi.org/10.1016/j.jmp.2019.102309

Müller, K. (2017). *here: A simpler way to find your files* (R package version 0.1) [Computer software manual]. https://CRAN.R-project.org/package=here

Nuthmann, A. (2017). Fixation durations in scene viewing: Modeling the effects of local image features, oculomotor parameters, and task. *Psychonom Review*, *117*(2), 382–405. https://doi.org/10.1037/a0018924

Nuthmann, A., Smith, T. J., Engbert, R., & Henderson, J. M. (2010). CRISP: A computational model of fixation durations in scene viewing. *Psychological Review*, *177*(2), 382–405. https://doi.org/10.1037/a0018924

Palmer, E. M., Horowitz, T. S., Torralba, A., & Wolfe, J. M. (2011). What are the shapes of response time distributions in visual search? *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 58–71. https://doi.org/10.1037/a0020747

Pedersen, T. L. (2019a). *ggforce: Accelerating 'ggplot2'* (R package version 0.3.1) [Computer software manual]. https://CRAN.R-project.org/package=ggforce

Pedersen, T. L. (2019b). *patchwork: The composer of plots* (R package version 1.0.0) [Computer software manual]. https://CRAN.R-project.org/package=patchwork

Polonio, L., Di Guida, S., & Coricelli, G. (2015). Strategic sophistication and attention in games: An eye-tracking study. *Games and Economic Behavior*, *94*, 80–96. https://doi.org/10.1016/j.geb.2015.09.003

Ratcliff, R., & McKoon, G. (2008). The di_usion decision model: theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

R Core Team. (2020). *R: A language and environment for statistical computing* [Computer software manual]. https://www.R-project.org/

Reichle, E. D., & Sheridan, H. (2015). EZ reader: An overview of the model and two recent applications. In A. Pollatsek & R. Treiman (Eds.), *The Oxford handbook of reading*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199324576.013.17

Robinaugh, D., Haslbeck, J., Ryan, O., Fried, E. I., & Waldorp, L. (2020). Invisible hands and _ne calipers: A call to use formal theory as a toolkit for theory construction. *PsyArXiv*. https://doi.org/10.31234/osf.io/ugz7y

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: a quiet methodological revolution. *American Psychologist*, *65*(1), 1–12.

Schad, D. J., Betancourt, M., & Vasishth, S. (2019). *Toward a principled Bayesian workow in cognitive science*. arXiv. https://arxiv.org/abs/1904.12765

Schütt, H. H., Rothkegel, L. O., Trukenbrod, H. A., Reich, S., Wichmann, F. A., & Engbert, R. (2017). Likelihood-based parameter estimation and comparison of dynamical cognitive models. *Psychological Review*, *124*(4), 505–524. https://doi.org/10.1037/rev0000068

Schwetlick, L., Rothkegel, L. O., & Engbert, R. (2019, September, 15). *Adding neurally-inspired mechanisms to the SceneWalk model improves scan path predictions for natural images* [paper presentation]. 2019 conference on cognitive computational neuro-science, Berlin, Germany. https://doi.org/10.32470/CCN.2019.1206-0

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, *27*(3), 379–423.

Smaldino, P. E. (2017). Models are stupid, and we need more of them. In R. R. Vallacher, S. J. Read, & A. Nowak (Eds.), *Computational social psychology* (pp. 311–331). Routledge.

Stan Development Team. (2020). *RStan: The R interface to Stan* (R package version 2.19.3) [Computer software manual]. http://mc-stan.org/

Tatler, B. W. (2007). The central _xation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*(14), Article 4. https://doi.org/10.1167/7.14.4

Tatler, B. W., Brockmole, J. R., & Carpenter, R. H. (2017). LATEST: A model of saccadic decisions in space and time. *Psychological Review*, *124*(3), 267–300. https://doi.org/10.1037/rev0000054

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision*, *11*(5), Article 5. https://doi.org/10.1167/11.5.5

Tatler, B. W., & Vincent, B. T. (2008). Systematic tendencies in scene viewing. *Journal of Eye Movement Research*, *2*(2), 1–18. https://doi.org/10.16910/jemr.2.2.5

Trukenbrod, H. A., & Engbert, R. (2014). ICAT: A computational model for the adaptive control of fixation durations. *Psychonomic Bulletin & Review*, *21*(4), 907–934. https://doi.org/10.3758/s13423-013-0575-0

Tseng, P.-H., Carmi, R., Cameron, I. G., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, *9*(7), Article 4. https://doi.org/10.1167/9.7.4

van Renswoude, D. R., Johnson, S., Raijmakers, M. E. J., & Visser, I. (2016). Do infants have the horizontal bias? *Infant Behavior and Development*, *44*, 38–48. https://doi.org/10.1016/j.infbeh.2016.05.005

van Renswoude, D. R., van den Berg, L., Raijmakers, M. E. J., & Visser, I. (2019). Infants' center bias in free viewing of real-world scenes. *Vision research*, *154*, 44–53. https://doi.org/10.1016/j.visres.2018.10.003

van Renswoude, D. R., Voorvaart, R. E., van den Berg, L., Raijmakers, M. E. J., & Visser, I. (2020). *Object familiarity inuences infant gaze control during free scene viewing* [Manuscript in preparation].

Wagenmakers, E.-J., Dutilh, G., & Sarafoglou, A. (2018). The creativity verification cycle in psychological science: New methods to combat old idols. *Perspectives on Psychological Science*, *13* (4), 418–427. https://doi.org/10.1177/1745691618771357

Wang, X.-S., & Wong, R. (2007). Discrete analogues of laplace's approximation. *Asymptotic Analysis*, *54* (3–4), 165–180.

Warnes, G. R., Bolker, B., & Lumley, T. (2020). *gtools: Various r programming tools* (R package version 3.8.2) [Computer software manual]. https://CRAN.R-project.org/package=gtools

Wickham, H. (2019). *tidyverse: Easily install and load the 'tidyverse'* (R package version 1.3.0) [Computer software manual]. https://CRAN.R-project.org/package=tidyverse

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., . . . Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Chapman and Hall/CRC. https://yihui.org/knitr/

Xie, Y. (2020). *knitr: A general-purpose package for dynamic report generationin r* (R package version 1.28) [Computer software manual]. https://CRAN.R-project.org/package=knitr

Xu, J., Jiang, M., Wang, S., Kankanhalli, M. S., & Zhao, Q. (2014). Predicting human gaze beyond pixels. *Journal of Vision*, *14*(1), Article 28. https://doi.org/10.1167/14.1.28

Yarkoni, T. (2019). The generalizability crisis. *PsyArXiv*. https://doi.org/10.31234/osf.io/jqw35

Zelinsky, G. J., Adeli, H., Peng, Y., & Samaras, D. (2013). Modelling eye movements in a categorical search task. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*(1628), 20130058. https://doi.org/10.1098/rstb.2013.0058

Received July 17, 2020
Revision received November 22, 2020
Accepted March 5, 2021 ■