BRAID-Color: Extending a probabilistic model of visual word recognition and reading with color processing to account for the Stroop effect

Bernardo Marques[1], **Camille Charrier**[2], **Alexandra Steinhilber**[2] **& Julien Diard**[2]
[1]Department of Philosophy, École Normale Supérieure, PSL, UMR 8241 République des Savoirs (CNRS/ENS/Collège de France), 29 rue d'Ulm, Paris, 75005, France
[2]Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LPNC, 1251 rue des Universités, Grenoble, 38000, France

Abstract

The Stroop effect is a well-known phenomenon where participants take longer when they name the color of letters that spell the name of a different color. This effect has been extensively studied, with various proposed theories and computational models. However, most models do not integrate with existing frameworks for color processing or reading, limiting their generalizability. We introduce the BRAID-Color probabilistic model, an extension of BRAID, an existing model of reading and reading acquisition. BRAID-Color consists in adding to an orthographic information processing branch, able to recognize words spelled by a sequence of letters, a color processing branch, able to name colored visual stimuli. Information fusion at the lexical level combines both to simulate color naming with colored letters. With default parameters, simulations in the congruent, neutral and incongruent conditions illustrate that the model successfully and robustly accounts for the Stroop effect.

Keywords: computational modeling; Stroop effect; probabilistic modeling; reading models

Introduction

The Stroop task (Stroop, 1935) is one of the most extensively studied phenomena in cognitive psychology. In the original experiment, participants were asked to name the color of letters, while the word they spelled was the name of a color. For instance, the word "green" might be presented in red ink, and the correct response would be "red". In the Stroop task, response times and error patterns suggest an interference between the two stimuli when they are not congruent: this is the Stroop effect. Despite almost a century of research since its introduction, a definitive explanation for the Stroop effect has yet to be established (Algom & Chajut, 2019; Palfi et al., 2022; Scarpina & Tagini, 2017).

Various methods have been employed to investigate the Stroop effect (Klein, 1964; MacLeod, 1991; Monsell et al., 2001; Tzelgov et al., 1992), with computational modeling taking a prominent role. The first notable model of the Stroop effect (Cohen et al., 1990) proposed a model of attention with two processing paths, for processing respectively color information and word information. The interference between the two paths was considered through task-control nodes. Their weights varied continuously, providing a hypothesis of increasing automaticity with task repetition. This model has been influential in the literature (Chuderski & Smolen, 2016; Cohen & Huston, 1994; Kanne et al., 1998).

Some connectionist models were generalizations from other models with broader scopes. For instance, Phaf et al. (1990) have proposed an application of SLAM (SeLective Attention Model), initially designed for visual selective attention tasks from McClelland and Rumelhart (1981) by adding components of response selection and evaluation. Another example is the connectionist model proposed by Roelofs (2000, 2003), an extension of the WEAVER++ model of word production (Levelt et al., 1999; Roelofs, 1992, 1997a, 1997b). This strategy was also employed by Altmann and Davidson (2001), who proposed the WACT model, resulting from the combination of WEAVER++ with the memory and control processes of the ACT-R cognitive theory (Anderson & Lebiere, 1998).

Computational models accounting for the Stroop effect should connect, ideally, both to color naming models, and to models of visual word recognition or reading, which are numerous. A useful taxonomy of models in this domain characterizes models according to the task they consider. Word recognition models (for reviews, see Norris, 2013; Reichle, 2021) are usually purely orthographic and consider letter and word recognition tasks, and some consider lexical decision tasks as well. Word reading models (Coltheart et al., 2001; Perry et al., 2007, 2010; Plaut et al., 1996; Seidenberg & McClelland, 1989) include orthographic and phonological representations and consider the task of reading aloud isolated words. Finally, text reading models (Engbert et al., 2002, 2005; Reichle et al., 1999, 2009) include eye-movement control mechanisms and consider sequences of words as stimuli, to simulate eye trajectories on the page. The resulting literature is fragmented, with separate families of models for each specific task or subdomain (Norris, 2013). A few exceptions are found, with models of reading extended to account for orthographic learning (Pritchard et al., 2018; Ziegler et al., 2014), a model of visual word recognition connected to text reading (Snell et al., 2018), a model of eye-movement control extended towards word recognition, lexical decision and sentence comprehension (Veldre et al., 2020), and finally, BRAID, a family of probabilistic models of letter recognitin, word recognition, lexical decision, reading and orthographic learning (Ginestet et al., 2019, 2022; Phénix et al., 2025; Steinhilber et al., 2023).

In this context, the contribution of Coltheart et al. (1999) is noteworthy. In this particular study, the Dual Route Cascaded (DRC) model of reading was extended to simulate color naming, with a particular focus on positional priming effects. The DRC model was modified by linking a semantic representa-

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

tion of colors to the phonological lexicon, and simulating the Stroop task involved an ad-hoc manipulation of the strength of connections in the letter-to-orthographic-lexicon network. To the best of our knowledge, this extension was absent of further presentations of the DRC model (e.g., Coltheart et al., 2001; Pritchard et al., 2018).

In summary, the computational models developed for studying the Stroop task have predominantly been task-specific (Botvinick et al., 2001; Chung et al., 2013; Kalanthroff et al., 2018; Siegle et al., 2004), that is to say, they were specifically developed to simulate the Stroop effect, without considering any other type of task related to reading, speech production or color perception. Two noteworthy exceptions include extending a computational model of speech production (Roelofs, 2003, 2010) and, more relevant to our approach, extending a computational model of reading (Coltheart et al., 1999).

In our eyes, a collection of task-specific cognitive models, however large it may be, does not constitute a solid cognitive theory, for two reasons. First, neurophysiological observation suggest that the central nervous system is not a collection of highly-specific processes (e.g., Roger et al., 2022). It seems indeed obvious that most of the visual system, for instance, is task agnostic, and involved for visual processing in various task contexts. Second, as a scientific program, developing computational models in isolation makes them largely unconstrained, resulting in a variety of mathematical frameworks, with inconsistent and sometimes incompatible predictions and theoretical interpretations (Norris, 2013; Veldre et al., 2020).

Therefore, in this paper, we aim at generalizing a prior model of reading to account for the Stroop effect, more than 25 years after the proposal by Coltheart et al. (1999). To this end, we developed BRAID-Color, an extension of the BRAID family of probabilistic models. To assess the ability of the BRAID family of models to generalize, we will study the BRAID-Color "as-is", that is to say without specific model or parameter adaptation, and verify whether this minimal extension can account for the Stroop effect.

The structure of our paper is as follows. We first introduce the BRAID family of reading models, and then BRAID-Color, its extension able to process the color of the stimulus. We then present the results of an experiment with simulations of the Stroop task with BRAID-Color, to assess its ability to account for the Stroop effect.

The BRAID-Color model

The BRAID family of models is a notable exception regarding the issue of task specificity. The BRAID models are hierarchical probabilistic models that aim to be general models of written language processing, irrespective of the specific task at hand. From the vanilla BRAID model (Ginestet et al., 2019; Phénix et al., 2025), that was able to simulate letter recognition, word recognition and lexical decision, its BRAID-Learn extension (Ginestet et al., 2022; Steinhilber et al., 2023) simulated orthographic learning, its BRAID-Phon extension (Saghiran et al., 2020) simulated reading aloud and finally, its BRAID-Acq extension (Steinhilber, 2023) simulated reading acquisition. These variants have been developed using a nested methodology, so that later variants are designed to also account for any behavioral effects in tasks simulated by earlier variants.

The BRAID-Color model mainly consists in adding a color processing branch to the existing architecture of BRAID family of models. However, we present it here as an extension of the vanilla BRAID model, which has fewer features to present. Therefore, the vanilla BRAID model is presented first, with a focus on features that are relevant to the current topic. Then, a description of its extension into BRAID-Color is provided, allowing for processing the color of the visual stimulus. The structure of the architecture of the BRAID-Color model is illustrated schematically in Figure 1.

The BRAID model

The structure of the BRAID model is illustrated on the right portion of Figure 1 (adapted with permission from Saghiran, 2021), and corresponds to the "orthographic branch" of the BRAID-Color model.

Its architecture is composed of four submodels. First, the letter sensory model (green box on the right of Figure 1) encodes the visual input as a sequence of letters, and includes three mechanisms to account for several features of low-level visual processing: decrease of performance as a function of eccentricity due to acuity, decrease of performance due to lateral interference from neighboring letters, and patterns of errors reflecting geometric similarity between letters.

Second, the letter perceptual submodel (blue box in Figure 1) receives the output of the letter sensory submodel as sensory evidence about the identity of letters in each position. This sensory evidence accumulates over time into probability distributions of Markov chains, one for each position within the letter sequence. Initially, these distributions are defined as uniform, representing a state of maximal uncertainty over letter identity. As the simulation progresses, the perceptual probability distributions peak on the correct letter identity (at least, in usual simulations with default parameters, where the visual stimulus is not significantly degraded or challenging).

Third, the visual attentional submodel (orange box in Figure 1) modulates the speed of information transfer between the letter sensory and letter perceptual submodels. This modulation affects the dynamics of sensory evidence accumulation into letter perceptual probability distributions. The spatial distribution of attention is represented, in the model, with a one-dimensional Normal probability distribution. Modifying its mean position and variance affects the dynamics of letter identification, consequently affecting the process of word recognition. Although not useful in the current context of simulating the Stroop effect, the visual attention submodel has been demonstrated to be effective in accounting for length effects in lexical decision (Ginestet et al., 2019), for crowding effects and the optimal viewing position effect (Phénix et al., 2025), and in depicting how visuo-attentional exploration pro-

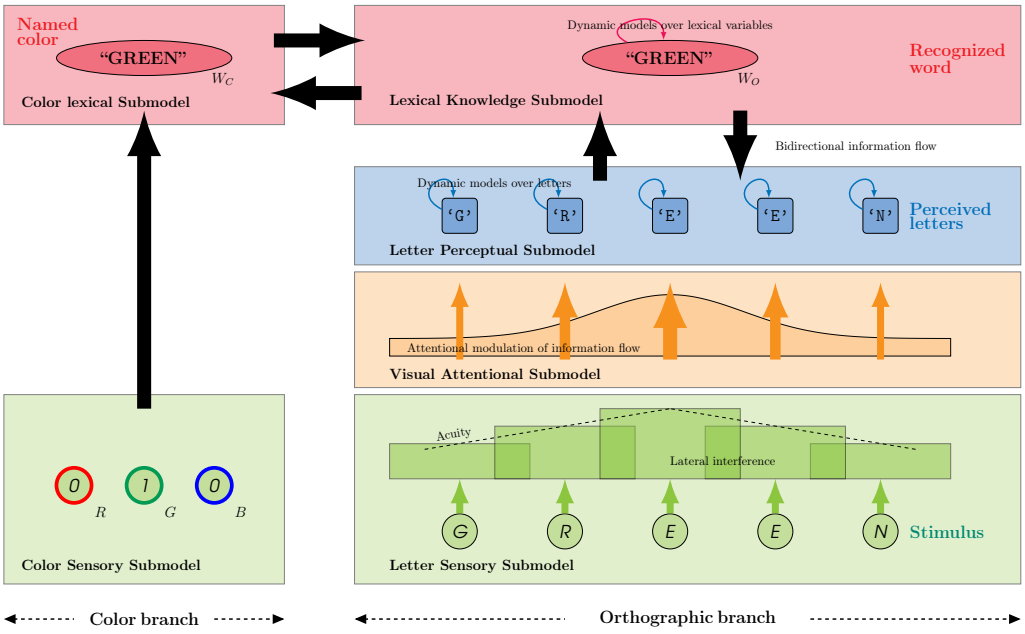

Figure 1: Graphical representation of the BRAID-Color model. Colored rectangles delineate submodels. The color processing branch of the model is on the left, the orthographic branch of the model (i.e., BRAID itself) on the right. This example illustrates a case of congruent stimulation, with the input letter sequence "GREEN" in the orthographic branch, which are colored in green ink (RGB coordinates are 0, 1, 0).

gresses from a slow and serial, letter-to-letter decoding to a fluent, parallel, global decoding during orthographic learning (Ginestet et al., 2022; Steinhilber et al., 2023).

Fourth and finally, the lexical knowledge submodel (red box in the orthographic branch of Figure 1) contains several components. First, the lexical space W is defined as the set of known words. Second, a Naïve Bayes probabilistic model encodes knowledge about which letters compose which words. Third, a Markov-chain model, mathematically similar to the perceptual accumulators of evidence of the letter perceptual submodel, provides accumulation of evidence concerning word identity. In other words, a dynamic probability distribution is computed at each simulated time-step over the set of all known words in the lexical space.

With these four submodels, the BRAID model can be used to simulate tasks such as letter perception, word recognition and lexical decision (not described here). Mathematically, each of these tasks is performed by applying the rules of Bayesian inference to the probabilistic definition of the model. Overall, these tasks proceed successfully in nominal conditions (Phénix et al., 2025): for instance, with default parameters, when the stimulus corresponds to a word in the lexical knowledge of the model, the probability distribution over the lexical space peaks on the correct word hypothesis.

Extending BRAID into BRAID-Color

Color processing submodel A fundamental assumption of the BRAID model is that its internal representation of letters only concerns identity and position, irrespective of any other characteristics of the stimulus, such as letter size, case, font, style (italic, bold, etc.), orientation or color. This is in line with neurophysiological studies (for a review, see Dehaene et al., 2005), that support such an hypothesis of an abstract cognitive representation of letter identity.

Nevertheless, visual representations in the central nervous system also encode information about the color, or colors, of visual stimuli. To define BRAID-Color, we define a model of color processing, that maps a sensory color space to a space of color names. Its mathematical form, and the resulting categorization process, are analogous to classical probabilistic models of perceptual categorization. For instance, it can be regarded as an application of a model of phonemic identification (Feldman et al., 2009a, 2009b; Kronrod et al., 2016; Vallabha et al., 2007), except that its sensory space is the three-dimensional (3D) color space instead of the acoustic formant space, and its categorization space is the set of color names instead of the set of known phonemes.

More precisely, the color space is defined as a 3D cube of Red-Green-Blue (RGB) values. Each R, G, B dimension value is a real number in the $[0, 1]$ interval. We consider that there are 8 named colors, represented by the discrete set

$$C = \{c_{red}, c_{green}, c_{blue}, c_{white}, c_{black}, c_{cyan}, c_{magenta}, c_{yellow}\} .$$

To map the color and color name spaces, we define the joint probability distribution as follows:

$$P(R\ G\ B\ C) = P(C)P(R\ G\ B \mid C),$$

where $P(C)$ is a uniform probability distribution (i.e., each color name has a prior probability of 1/8) and $P(R\ G\ B \mid C)$

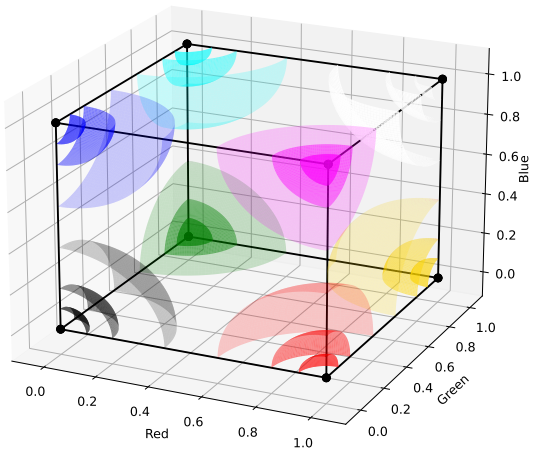

Figure 2: Illustration of the probabilistic model $P(R\ G\ B \mid C)$ for color perception. Axes represent the three color coordinates for red, green and blue. Three iso-probability surfaces for each 3D Normal probability distribution are shown, and are colored to match the corresponding color name.

is a collection of eight 3D Normal probability distributions. These are defined by mean vector values and covariance matrices, which map regions of the color space to each color name. There are eight mean vector values, denoted by $\mu_R(C), \mu_G(C), \mu_B(C)$, one for each color. For instance:

$$\begin{array}{lll} \mu_R(c_{red}) = 1, & \mu_G(c_{red}) = 0, & \mu_B(c_{red}) = 0; \\ \mu_R(c_{green}) = 0, & \mu_G(c_{green}) = 1, & \mu_B(c_{green}) = 0; \end{array}$$

and so on, with each mean position at each corner of the color cube. The covariance matrices are independent of the color name, and are all equal to a diagonal covariance matrix $\Sigma_{RGB} = \sigma^2.I$, with σ^2 the variance parameter and I the 3*3 identity matrix (i.e., we assume equal dispersion along each color dimension, that is, spherical iso-probability surfaces). Variance σ^2 was set to a high value so that probability distributions over colors would retain high uncertainty, to be adequately scaled with other distributions in the model. (More precisely, σ^2 was arbitrarily set to 400, with the internal coding of the model based on a [0, 255] RGB space, i.e., $\sigma^2 = 400/255 = 1.57$ in the current paper's chosen description.) This is illustrated in Figure 2.

From the joint probability distribution $P(R\ G\ B\ C)$, we obtain a categorization process, by computing, from any given color stimulus r, g, b, a posterior probability distribution over the color name space C. This is noted $P(C \mid r\ g\ b)$. Bayesian inference yields:

$$P(C \mid r\ g\ b) \quad = \quad \frac{P(r\ g\ b \mid C)}{\sum_C P(r\ g\ b \mid C)}\ . \qquad (1)$$

Including color processing into BRAID In order to incorporate our proposed model of color processing into the BRAID model, it must be noted that the eight color names of C are a subset of the lexical space W of known words of BRAID. In the following simulations, and for simplicity, the lexical space is constrained to fixed-length words, retaining only five-letter words. Therefore, color names that are not five-letter-long are internally coded with word-like pseudo-words (to ensure they have lexical neighbors without any particularity) as follows: "reder" (neighbors such as "rider", "ruder" and "elder"), "blues" (neighbors such as "clues" and "glues"), "cyane" (neighbors such as "crane" and "plane"), "maven" (neighbors such as "haven", "given" and "riven") and "yello" (neighbors such as "hello" and "cello"). This has no impact on the results or their ability to generalize.

To connect the lexical space to the color space C, there are two cases to consider: for the eight color names in W that also correspond to categories in C, their relation with the sensory color space is characterized by the 3D Normal probability distributions described above; for all remaining words in W, we assume 3D uniform probability distributions over the color space R, G, B. In the mathematical definition of the color processing branch BRAID-Color, we denote W_C the variable whose domain is the complete lexical space and that is connected to the R, G, B color space; instead of Eq. (1), we compute $P(W_C \mid r\ g\ b)$. In summary, in the context of BRAID-Color, there are only eight color names in the lexical space that are informed by color processing. For instance, names of other colors (e.g., "mauve") or names of objects with typical colors (e.g., "banana") have no particular association to colors. This is of course a simplifying assumption.

Consider the connection between the color and orthographic branches of processing: in Figure 1, it is represented by bidirectional arrows. Mathematically, this is implemented with a "controlled coherence variable" noted λ_C, which is controlled by parameter α (Ginestet et al., 2019; Nabé et al., 2022). Coherence variables are defined as a set of variables that allow connecting variables sharing the same mathematical domain. They yield multiplicative models of probabilistic fusion of information. The controlled version of coherence variables, through their control parameter, enables the modulation of the relative information content of the connected probability distributions during their combination.

In the BRAID-Color model, the result of color sensory processing is thus controlled with respect to its information content, as a function of parameter α: what is combined with the rest of the model is not $P(W_C \mid r\ g\ b)$, but

$$\alpha P(W_C \mid r\ g\ b) + (1 - \alpha)\mathbf{U},$$

with $\mathbf{U}$ the uniform probability distribution over the lexical space. This weighted sum between $P(W_C \mid r\ g\ b)$ and a uniform distribution allows to gradually go from the distribution itself to the uniform, as a function of α. Since the uniform distribution is devoid of any information, this operation allows modulating "how much" $P(W_C \mid r\ g\ b)$ is considered during fusion. This is mathematically useful for controlling the relative balance of information content between the color and orthographic branches of the BRAID-Color model. Indeed, as it involves well-separated 3D Normal distributions, color processing can result in probability distributions with low entropy, that is, almost all-or-nothing distributions with

numerical values close to ones and zeroes. This particularly happens when color stimulation in the RGB color space approaches one of the corners of the RGB cube. In such circumstances, the term $P(W_C \mid r\ g\ b)$ has low entropy. Consequently, the color branch prevails over any processing that occurs in the orthographic branch. We will therefore study the impact of parameter α on the capacity of the model to simulate the Stroop effect. To that end, four empirically set values of α will be considered: .1, .25, .5 and .75.

Simulating the Stroop task in BRAID-Color

In the BRAID-Color model, three distinct tasks are considered. The first task is classical word recognition. In this task, an input letter string $s^{1:T}_{1:N}$ is presented as stimulus, in positions 1 to N and from time instant 1 to T. The model computes the evolution of the probability distribution over the lexical space W, for each time instant T. Mathematically, solving word recognition is noted

$$Q^T_W = P(W^T_O \mid s^{1:T}_{1:N}) \ .$$

The resulting Bayesian inference is provided elsewhere (Phénix et al., 2025). Note however that we use here a simplified notation. In existing BRAID presentations (Phénix et al., 2018, 2025), gaze position $g^{1:T}$ and visuo-attentional parameters $\mu^{1:T}_A$ and $\sigma^{1:T}_A$ are specified in Q^T_W, along with controlled coherence variable values to control information flow in the model. Since gaze and visual attention are not manipulated in the following experiments, and to keep mathematical notation light, we omit these. With recognition defined as Q^T_W, we remark that the color stimulation is not even specified in this computation (variables R, G, B do not appear in Q^T_W). Computing Q^T_W with Bayesian inference involves a marginalization over the RGB color space, which simplifies out. In other words, when the color of the stimulus is not specified, the color submodel has no impact whatsoever, and word recognition proceeds as in the BRAID model.

The second task we consider is color naming, which can be regarded as a mirror situation to the previous task. The stimulus is simply a color stimulation r, g, b, and no letters are shown as stimulus, as if, for instance, a color patch was presented on a screen. This is mathematically denoted as Q^T_C. Bayesian inference, thanks to the naïve Bayes structure of information accumulating in the lexical space, also simplifies out. The resulting Bayesian inference was previously given:

$$Q^T_C = \alpha P(W_C \mid r\ g\ b) + (1-\alpha)\mathbf{U} \ .$$

The third and final task we consider, of course, is the Stroop task, that is, naming the color of letters that spell out the name of a color. This is noted:

$$Q^T_{CW} = P(W^T \mid r\ g\ b\ s^{1:T}_{1:N}\ [\lambda_C = 1]) \ .$$

Setting the controlled coherence variable λ_C to value 1 "closes the Bayesian switch" (Gilet et al., 2011) between the color and orthographic branches of the model, allowing exchange of information between them. Mathematically, Bayesian inference to compute Q^T_{CW} is equivalent to a simple product combination of Q^T_W and Q^T_C:

$$\begin{aligned} Q^T_{CW} &\propto Q^T_W * Q^T_C \\ &\propto P(W^T \mid s^{1:T}_{1:N}) * \left[\alpha P(W_C \mid r\ g\ b) + (1-\alpha)\mathbf{U}\right] , \end{aligned} \tag{2}$$

with $\propto$ the proportionality relation (in other words, the product $Q^T_W * Q^T_C$ is renormalized over the lexical space W^T).

Experiments and results

The Stroop effect, in its simplest form, consists in a facilitation of color naming by word recognition in congruent conditions in comparison to neutral condition, and, conversely, an interference of word recognition on color naming in incongruent conditions compared to neutral conditions. We therefore consider three experimental conditions.

In the incongruent condition, the sequence of letters in the stimulus spells out the name of a different color than the color of the ink in which it is presented. Given that the model represents eight colors, all $8 * 7 = 56$ possible combinations are considered (eight possible ink colors times seven letter sequences that spell another color name). Each of these combinations is then used as stimulus for the Stroop task. In the neutral condition, the stimulus letter sequence corresponds to a word that has no particular relation to color names. For each of the eight possible ink colors provided as color stimulation, we randomly selected seven words in the lexicon, and provided them as stimulus letter sequences. This also results in 56 different simulations. Finally, in the congruent condition, the stimulus letter sequence spells out the name of the ink color in which it is presented; this only results in eight possible combinations.

Mathematically, in the BRAID-Color model, these are simulated by three kinds of instances of the Stroop task. For instance: $Q^T_{CW} = P(W^T \mid [r = 1]\ [g = 0]\ [b = 0]\ [s^{1:T}_{1:5} = \textit{"GREEN"}])$ is an example of the incongruent condition with the word "GREEN" presented in red ink; $Q^T_{CW} = P(W^T \mid [r = 1]\ [g = 0]\ [b = 0]\ [s^{1:T}_{1:5} = \textit{"HOUSE"}])$ is an example of the neutral condition with the word "HOUSE" presented in red ink; and $Q^T_{CW} = P(W^T \mid [r = 0]\ [g = 1]\ [b = 0]\ [s^{1:T}_{1:5} = \textit{"GREEN"}])$ is an example of the congruent condition with the word "GREEN" presented in green ink.

We now present simulation results. Consider first the correctness of responses: in all simulations, the probability distribution over the lexical space computed with Q^T_{CW} assigns its highest probability value to the correct color name. Consider now the dynamics of perceptual processing in the model, illustrated Figure 3. We observe faster increase in probability for the correct color name in lexical space for the congruent condition relative to the neutral condition, and for the neutral condition relative to the incongruent condition. The separation between the dynamics of each condition is larger for larger values of α which, we recall, represent a stronger influence of color processing relative to letter processing.

In order to obtain either response times (RT) or correct response rates from these probability distribution evolution,

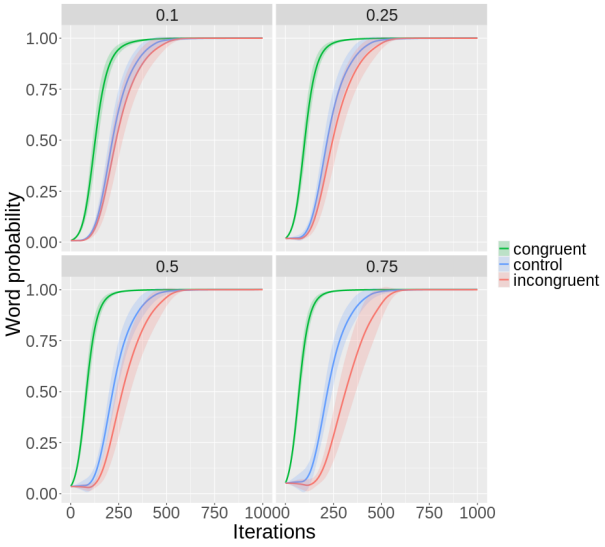

Figure 3: Simulation of the Stroop task. Graphs show the evolution of the probability value of the most probable word (y-axis), as a function of simulated time (x-axis). Each solid curve shows the evolution of average word probability over all stimuli of the congruent condition (green), neutral condition (blue, labeled "control") or incongruent condition (red). Variation ribbons around the average response curves indicate variance in response curves (recall that the neutral and incongruent conditions feature 56 simulations, whereas the congruent condition only has 8, resulting in reduced variance). Each panel corresponds to a different value of parameter α (.1, .25, .5 and .75), with higher values representing larger influence of color processing in the model.

decision methods would need to be defined. Concerning response rates, and since probability distributions quickly peak towards the correct color name in the lexical space, any decision model selecting the most likely hypothesis in the lexical space would yield 100 % correct response rates. Decision models based on random drawing according to Q^T_{CW} would also provide high correct response rates, and the higher the longer the stimulation time.

Consider now response times. A classical decision model would assume that reaching a given probability threshold, denoted τ, would stop perceptual accumulation and elicit a response. On Figure 3, this would amount to reading time instants when probability values reach some horizontal line representing probability τ. It is easily seen in Figure 3 that, for a wide range of such decision thresholds, and for all values of parameter α, the predicted RTs would be smaller in the congruent condition relative to the neutral condition, and the predicted RTs would be smaller in the neutral condition relative to the incongruent condition. Therefore, the BRAID-Color model simulates a facilitation for the congruent condition, and an interference for the incongruent condition, in a robust fashion; it is thus able to account for the Stroop effect.

Discussion

With the exception of a single study in the connectionist framework (Coltheart et al., 1999), to the best of our knowledge, the investigation of whether a prior computational model of reading could be extended to simulate the Stroop effect has remained largely unexplored in psychological research. We proposed to fill this gap by examining the potential of a reading model to be minimally extended to account for the Stroop effect. In this sense, our objective differed from previous efforts focused on the development of task-specific models of the Stroop task. These focused on accounting for the Stroop effect and its variants in precise ways. Instead, our objective was to explore the generalization of an existing reading model to the color naming task and the Stroop effect.

Consequently, we have introduced the BRAID-Color model, an extension of the BRAID model of visual word recognition. By connecting a submodel of visual color processing to the existing orthographical architecture, we have shown the model's capability to simulate not only color naming in general but also the naming of the color of letters that spell a word. In this last case, the mathematical properties of the model yield a combination of probability distributions at the lexical level. This combination results in an interference between lexical information from the letter sequence and from letter color when they are incongruent, and, in contrast, a facilitation when both information sources are congruent. As a consequence of this mathematical property, we have demonstrated in our simulations that the BRAID-Color model would produce the expected Stroop effect.

This shows that the BRAID model can be generalized to color processing, allowing the simulation of the color naming task and accounting for the Stroop effect. This mathematically straightforward extension involves the addition of a submodel featuring the obviously missing component, namely, color representations and their relations to the lexical space. More importantly, our study did not involve any change of the parameters of the BRAID model. Nevertheless, the mechanism of information combination at the lexical level robustly produces the Stroop effect.

Of course, such a straightforward extension, without any parameter adaptation or calibration, and with simple "sanity-check" simulations, could not be expected to account for the Stroop effect in all its richness. For instance, a salient feature of behavioral accounts of the Stroop effect is the presence of an asymmetry, whereby a larger interference is observed in response to incongruent stimuli relative to the facilitation effect elicited by congruent stimuli (Chuderski & Smolen, 2016). Our experimental results suggest that the model, with default parameters, produces the inverse pattern, with larger facilitation than interference (see Figure 3). Parameter tuning, maybe limited to a more realistic tuning of the dynamics of color recognition, would affect this aspect of our experimental results, and likely bring the model closer to realistic simulations.

Acknowledgments

Authors wish to thank Sylviane Valdois and Marie-Line Bosse for countless conversations and support.

References

Algom, D., & Chajut, E. (2019). Reclaiming the Stroop effect back from control to input-driven attention and perception. *Frontiers in Psychology*, *10*, 1–17.

Altmann, E., & Davidson, D. (2001). An integrative approach to Stroop: Combining a language model and a unified cognitive theory. *Proceedings of the 23rd annual meeting of the Cognitive Science Society*, 21–26.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Lawrence Erlbaum Associates, Inc.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychol. Rev.*, *108 (3)*, 624–652.

Chuderski, A., & Smolen, T. (2016). An integrated utility-based model of conflict evaluation and resolution in the Stroop task. *Psychological Review*, *123*, 255–290.

Chung, D., Raz, A., Lee, J., & Jeong, J. (2013). Computational modeling of the negative priming effect based on inhibition patterns and working memory. *Frontiers in Computational Neuroscience*, *7*.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the stroop effect. *Psychological Review*, *97 (3)*, 332–361.

Cohen, J. D., & Huston, T. A. (1994). Progress in the use of interactive models for understanding attention and performance. The MIT Press.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. C. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*(1), 204–256.

Coltheart, M., Woollams, A., Kinoshita, S., & Perry, C. (1999). A position-sensitive Stroop effect: Further evidence from a left-to-right component in print-to-speech conversion. *Psychonomic Bulletin & Review*, *6*(3), 456–463.

Dehaene, S., Cohen, L., Sigman, M., & Vinckier, F. (2005). The neural code for written words: A proposal. *Trends in Cognitive Sciences*, *9*(7), 335–341.

Engbert, R., Longtin, A., & Kliegl, R. (2002). A dynamical model of saccade generation in reading based on spatially distributed lexical processing. *Vision Research*, *42*, 621–636.

Engbert, R., Nuthmann, A., Richter, E. M., & Kliegl, R. (2005). SWIFT: A dynamical model of saccade generation during reading. *Psychological Review*, *112*(4), 777–813.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009a). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*(4), 752–782.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009b). Learning phonetic categories by learning a lexicon. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, 2208–2213.

Gilet, E., Diard, J., & Bessière, P. (2011). Bayesian action-perception computational model: Interaction of production and recognition of cursive letters. *PLoS ONE*, *6*(6), e20387.

Ginestet, E., Phénix, T., Diard, J., & Valdois, S. (2019). Modeling the length effect for words in lexical decision: The role of visual attention. *Vision Research*, *159*, 10–20.

Ginestet, E., Valdois, S., & Diard, J. (2022). Probabilistic modeling of orthographic learning based on visuo-attentional dynamics. *Psychonomic Bulletin & Review*, *29*, 1649–1672.

Kalanthroff, E., Davelaar, E., Henik, A., Goldfarb, L., & Usher, M. (2018). Task conflict and proactive control: A computational theory of the Stroop task. *Psychological Review*, *125 (1)*, 59–82.

Kanne, S. M., Balota, D. A., Spieler, D. H., & Faust, M. E. (1998). Explorations of cohen, dunbar, and mcclelland's (1990) connectionist model of Stroop performance. *Psychological Review*, *105 (1)*, 174–187.

Klein, G. S. (1964). Semantic power measured through the interference of words with color-naming. *The American Journal of Psychology*, *77 (No. 4)*, 576–88.

Kronrod, Y., Coppess, E., & Feldman, N. H. (2016). A unified account of categorical effects in phonetic perception. *Psychonomic Bulletin & Review*, *23*, 1681–1712.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22 (1)*, 1–38.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109 (2)*, 163–203.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. an account of basic findings. *Psychological Review*, *88 (5)*, 375–407.

Monsell, S., Taylor, T. J., & Murphy, K. (2001). Naming the color of a word: Is it responses or task sets that compete? *Memory & Cognition*, *29 (1)*, 137–151.

Nabé, M., Schwartz, J.-L., & Diard, J. (2022). Bayesian gates: A probabilistic modeling tool for temporal segmentation of sensory streams into sequences of perceptual accumulators. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the 44th annual conference of the cognitive science society* (pp. 2257–2263).

Norris, D. (2013). Models of visual word recognition. *Trends in Cognitive Sciences*, *17*(10), 517–524.

Palfi, B., Parris, B., Collins, A., & Dienes, Z. (2022). Strategies that reduce Stroop interference. *R. Soc. Open Sci.*, *9 (3)*, 163–203.

Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: The CDP+ model of reading aloud. *Psychological Review*, *114*(2), 273–315.

Perry, C., Ziegler, J. C., & Zorzi, M. (2010). Beyond single syllables: Large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. *Cognitive Psychology*, *61*(2), 106–51.

Phaf, R. H., Van der Heijden, A. H., & Hudson, P. T. (1990). Slam: A connectionist model for attention in visual selection tasks. *Cognitive Psychology*, *22 (3)*, 273–341.

Phénix, T., Ginestet, É., Valdois, S., & Diard, J. (2025). Visual attention matters during word recognition: A Bayesian modeling approach. *Psychonomic Bulletin & Review*.

Phénix, T., Valdois, S., & Diard, J. (2018). Reconciling opposite neighborhood frequency effects in lexical decision: Evidence from a novel probabilistic model of visual word recognition. In T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th annual conference of the cognitive science society* (pp. 2238–2243). Cognitive Science Society.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*(1), 56–115.

Pritchard, S. C., Coltheart, M., Marinus, E., & Castles, A. (2018). A computational model of the self-teaching hypothesis based on the dual-route cascaded model of reading. *Cognitive Science*, 1–49.

Reichle, E. D. (2021). *Computational models of reading: A handbook*. Oxford University Press.

Reichle, E. D., Rayner, K., & Pollatsek, A. (1999). Eye movement control in reading: Accounting for initial fixation locations and refixations within the E-Z Reader model. *Vision Research*, *39*, 4403–4411.

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using E-Z Reader to model the effects of higher-level language processing on eye movements during reading. *Psychonomic Bulletin & Review*, *16*(1), 1–21.

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition*, *42 (1)*, 107–142.

Roelofs, A. (1997a). Syllabification in speech production: Evaluation of weaver. *Language and Cognitive Processes*, *12*, 657–693.

Roelofs, A. (1997b). The WEAVER model of word-form encoding in speech production. *Cognition*, *64*, 249–284.

Roelofs, A. (2000). Control of language: A computational account of the Stroop asymmetry. In N. Taatgen & J. Aasman (Eds.), *Proceedings of the third international conference on cognitive modeling* (pp. 234–241).

Roelofs, A. (2003). Goal-referenced selection of verbal action: Modeling attentional control in the Stroop task. *Psychological Review*, *110 (1)*, 88–125.

Roelofs, A. (2010). Attention and facilitation: Converging information versus inadvertent reading in Stroop task performance. *J Exp Psychol Learn Mem Cogn*, *36 (2)*, 411–422.

Roger, E., Rodrigues De Almeida, L., Loevenbruck, H., Perrone-Bertolotti, M., Cousin, E., Schwartz, J.-L., Perrier, P., Dohen, M., Vilain, A., Baraduc, P., Achard, S., & Baciu, M. (2022). Unraveling the functional attributes of the language connectome: Crucial subnetworks, flexibility and variability. *NeuroImage*, *263*, 119672.

Saghiran, A. (2021). *Modélisation bayésienne de la lecture* [Doctoral dissertation, Univ. Grenoble Alpes].

Saghiran, A., Valdois, S., & Diard, J. (2020). Simulating length and frequency effects across multiple tasks with the Bayesian model BRAID-Phon. *Proceedings of the 42th Annual Conference of the Cognitive Science Society*, 3158–3163.

Scarpina, F., & Tagini, S. (2017). The Stroop color and word test. *Front Psychol.*, *8*, 557.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523–568.

Siegle, G., Steinhauer, S., & Thase, M. (2004). Pupillary assessment and computational modeling of the Stroop task in depression. *J Psychophysiol*, *52 (1)*, 63–76.

Snell, J., van Leipsig, S., Grainger, J., & Meeter, M. (2018). OB1-Reader: A model of word recognition and eye movements in text reading. *Psychological Review*, *125*(6), 969–984.

Steinhilber, A. (2023). *Modélisation bayésienne de l'apprentissage de la lecture* [Doctoral dissertation, Univ. Grenoble Alpes].

Steinhilber, A., Diard, J., Ginestet, E., & Valdois, S. (2023). Visual attention modulates the transition from fine-grained, serial processing to coarser-grained, more parallel processing: A computational modeling study. *Vision Research*, *207*, 108211, 1–15.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18 (6)*, 643–662.

Tzelgov, J., Henik, A., & Berger, J. (1992). Controlling Stroop effects by manipulating expectations for color words. *Mem Cogn*, *20*, 727–735.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, *104*(33), 13273–13278.

Veldre, A., Yu, L., Andrews, S., & Reichle, E. D. (2020). Towards a complete model of reading: Simulating lexical decision, word naming and sentence reading with Über-Reader. *Proceedings of the 42th Annual Conference of the Cognitive Science Society*.

Ziegler, J. C., Perry, C., & Zorzi, M. (2014). Modelling reading development through phonological decoding and self-teaching: Implications for dyslexia. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *369*(1634), 20120397.