A Unifying Model of Grapheme-Color Associations in Synesthetes and Controls

Nicholas B. Root (n.root@uva.nl)
Department of Psychology, University of Amsterdam, Roetersstraat 15,
Amsterdam, 1018WB, Noord-Holland, Netherlands

Romke Rouw (R.Rouw@uva.nl)
Department of Psychology, University of Amsterdam, Roetersstraat 15,
Amsterdam, 1018WB, Noord-Holland, Netherlands

Abstract

Grapheme-color synesthetes experience linguistic symbols as having a consistent color (e.g., "The letter R is burgundy."). Intriguingly, certain letters tend to be associated with certain colors, and these biases are not random: numerous properties of letters influence *which* letter is associated with *which* color. These influences, called "Regulatory Factors" (RFs), each explain some fraction of the variation in observed associations. No comprehensive model of the influences on grapheme-color associations exists: RFs have only been measured in isolation, are not always operationalized consistently, and often make competing predictions that cannot be accounted for in a univariate model. Here, we describe a statistical framework that integrates the predictions of all RFs into a single model, and thus yields a unified account of their influence on grapheme-color associations. Our model also links these predictions to measurable properties of language, offering a window into the multifactorial contributions to letter representation in the brain.

Keywords: synesthesia, predictive model, grapheme-color association, letter representation

Introduction

Grapheme-color synesthetes experience linguistic symbols (e.g., letters of the alphabet) as having a consistent color (e.g., "The letter R is burgundy"). Synesthesia is not a "disease" (indeed, synesthetes typically find their experiences pleasant and useful); it is instead a perceptual phenomenon with objectively measurable behavioral, electrophysiological, and neural characteristics (for a review, see Lovelace, 2013).

Synesthetes' specific grapheme-color associations (*which* letter is *which* color) are highly consistent over time: a synesthete will pick the same color for each grapheme even when re-tested after months or years (Asher et al., 2006). In addition, although the associations of any two synesthetes often differ (for one synesthete, "R is burgundy"; for another, "R is lime green"), studies of the associations of large groups of synesthetes reveals systematic biases in grapheme-color associations: certain graphemes are more frequently associated with certain colors (e.g. Simner et al., 2005; for a review see Simner, 2013). Intriguingly, similar biases in grapheme-color associations are observed if *non-synesthetes* are forced to choose colors for letters (Simner et al., 2005), suggesting that these biases are not caused by synesthesia *per se*, but might instead reflect underlying processes in all of us.

Why are certain letters likelier to be associated with certain colors? Numerous studies have reported correlations between synesthetic colors and various properties of letters, such as ordinal position, frequency in the language, and even pronunciation (for a review, see Simner, 2013). These influences, called **"Regulatory Factors"** (**RFs**; Asano & Yokosawa 2013; Root et al., *In Review*), each explain some fraction of the variation in observed associations.

Understanding why certain grapheme-color combinations are more frequent than others has implications beyond the field of synesthesia research: the fact that many RFs are linguistic in nature suggests that grapheme-color associations reflect language representation in the brain, and the fact that non-synesthetes seem to share similar (but non-conscious) grapheme-color associations suggests that the difference between synesthetes and non-synesthetes is in their conscious experience of color, rather than in their underlying letter representations. Synesthetes' conscious, specific, consistent associations are very easy to measure, and thus offer an extraordinary opportunity to study these representations.

Most published research on synesthetic RFs has examined only one RF, and different RFs are often studied by different researchers. As a result, RFs have been operationalized and analyzed in very different ways, and this methodological variation makes it difficult or impossible to quantitatively compare RFs. For example, the "Color Term" RF (e.g., Simner et al., 2005) predicts that color terms influence the colors associated with the initial letters ("Y is yellow", etc.), whereas the "Letter Frequency" RF (e.g., Beeli et al., 2007) predicts that more frequent letters are associated with brighter, more saturated colors. The Color Term RF yields an effect size measure of risk ratio in a categorical color space (e.g., $P(yellow \mid \text{"Y"})/P(yellow \mid !\text{"Y"})$), whereas the Letter Frequency RF yields an effect size measure of r^2 in a continuous color space. Furthermore, univariate estimates of RF effect size may be confounded by other RFs. For example, the Color Term RF predicts that "Y" is associated with yellow (a bright color), whereas the Letter Frequency RF predicts that "Y" (a low-frequency letter) is associated with a dark color. If both RFs "compete" to influence the color of "Y", a multivariate model that can account for the interaction between the RFs will yield estimates of RF effect size that are more accurate than the estimates of a univariate model.

In the present work, we describe a novel unified model of synesthetic RFs. In our framework, the predictions of all RFs are transformed into the same color space, and all RFs are modeled simultaneously, yielding effect size estimates that can be quantitatively compared. We demonstrate how our

model can be used to compare RF effect sizes across different operationalizations of the same RF, across different RFs, and even across synesthetes with different native languages.

A Unifying Model of Synesthetic RFs

Below, we describe a unifying model of synesthetic RFs that outputs the predicted distribution of color associations for each grapheme. We construct our model as follows.

Model Inputs

We are given:

- A set $\mathbf{G}$ consisting of a finite number $N_{\mathbf{G}}$ of distinct graphemes $\mathbf{G} = \{G_1, G_2, G_3, \ldots, G_{N_{\mathbf{G}}}\}$
- A set $\mathbf{C}$ consisting of a finite number $N_{\mathbf{C}}$ of "colors" $\mathbf{C} = \{C_1, C_2, C_3, \ldots, C_{N_{\mathbf{C}}}\}$, which are nonempty pairwise disjoint subsets of color space.
- A set $\mathbf{S}$ consisting of a finite number $N_{\mathbf{S}}$ of distinct test subjects $\mathbf{S} = \{S_1, S_2, S_3, \ldots, S_{N_{\mathbf{S}}}\}$.
- A dataset $\mathbf{X}$ consisting of a finite number $N_{\mathbf{X}}$ of "associations" $\boldsymbol{X} = \{X_1, X_2, X_3, \ldots, X_{N_X}\}$, each of which is an ordered triple of the form $X_i = (g_1, c_i, s_i)$ where $g_i \in \mathbf{G}$, $c_i \in \mathbf{C}$, and $s_i \in \mathbf{S}$. Note that $N_{\mathbf{X}} \leq N_{\mathbf{S}} N_{\mathbf{G}}$: each subject associates colors with at most $N_{\mathbf{G}}$ graphemes, but can also associate no color for some grapheme(s).
- A set $\mathbf{K}$ consisting of a finite number $N_{\mathbf{K}}$ of RFs ("regulatory factors") $\mathbf{K} = \{K_1, K_2, K_3, \ldots, K_{N_{\mathbf{K}}}\}$
- For each RF $k \in \mathbf{K}$, grapheme $g \in \mathbf{G}$, and color $c \in \mathbf{C}$, and subject $s \in \mathbf{S}$: the predicted probability $\pi_{k,g,c,s}$ that subject s associates grapheme g is with color c if the color of g is due entirely to the influence of RF k. All such probabilities are derived theoretically or determined by observation/experiment using subjects not in $\mathbf{S}$. Note that $\sum_{c \in \mathbf{C}} \pi_{k,g,c,s}$ need not equal 1, because not all RFs cause the association of some color with every grapheme for every subject.

Model Definition

From the above, it follows that for each RF $k \in \mathbf{K}$, grapheme $g \in \mathbf{G}$, and subject $s \in \mathbf{S}$, the probability $\gamma_{k,g,s}$ that subject s associates grapheme g with *any* color $c \in \mathbf{C}$ if the color of g is due entirely to the influence of RF k is given by:

$$\gamma_{k,g,s} = \sum_{c \in \mathbf{C}} \pi_{k,g,c,s} \tag{1}$$

Likewise, it follows that for each RF $k \in \mathbf{K}$, grapheme $g \in \mathbf{G}$, color $c \in \mathbf{C}$, and subject $s \in \mathbf{S}$, the conditional probability $\theta_{k,g,c,s}$ that subject s associates grapheme g with a particular color c – given that the color of the grapheme g is due entirely to the influence of RF k – is given by:

$$\theta_{k,g,c,s} = \begin{cases} \dfrac{\pi_{k,g,c,s}}{\gamma_{k,g,s}} & if\ \gamma_{k,g,s} \neq 0 \\ 0 & otherwise \end{cases} \tag{2}$$

$$\text{so that } \pi_{k,g,c,s} = \gamma_{k,g,s} \theta_{k,g,c,s}$$

$$\left(note: \sum_{c \in \mathbf{C}} \theta_{k,g,c,s} = 1 \right) \tag{2}$$

We associate a "relative strength" $\beta_i \geq 0$ with each RF $K_i \in \mathbf{K}$ (e.g., $\beta_3 = 2$ and $\beta_5 = 6$ would imply that if $\gamma_{k=3,g,s} = \gamma_{k=5,g,s}$ for all graphemes $g \in \mathbf{G}$ and subjects $s \in \mathbf{S}$, RF K_5 is three times more influential in causing color associations than RF K_3); and then define, for each RF $k \in \mathbf{K}$, each grapheme $g \in \mathbf{G}$, each subject $s \in \mathbf{S}$, and each vector $\vec{\beta} = \langle \beta_1, \beta_2, \beta_3, \ldots, \beta_{N_{\mathbf{K}}} \rangle$ of relative RF strengths:

$$\alpha_{k,g,s}(\vec{\beta}) = \frac{\beta_k \gamma_{k,g,s}}{\sum_{k \in \mathbf{K}} \beta_k \gamma_{k,g,s}} \tag{3}$$

$$\left(\text{note: } \sum_{k \in \mathbf{K}} \alpha_{k,g,s}(\vec{\beta}) = 1 \right)$$

We can now model the influence of regulatory factors on grapheme-color associations as a classical mixture model of categorical probability distributions θ, with $\vec{\beta}$ and γ as hyperparameters on the mixture weights α. That is, for any grapheme $g \in \mathbf{G}$ and subject $s \in S$, the probability that subject s will associate grapheme g with color $c \in \mathbf{C}$ is:

$$\sum_{k \in \mathbf{K}} \alpha_{k,g,s}(\vec{\beta}) \theta_{k,g,c,s} \tag{4}$$

This model is called a "mixture model" because it considers the dataset $\mathbf{X}$ of grapheme-color associations to be a "mixture" (union) of disjoint subgroups, one subgroup for each of the $N_{\mathbf{K}}$ different possible values of $k \in \mathbf{K}$, the (unobservable) latent variable "regulatory factor". The "mixture weight" $\alpha_{k,g,s}(\vec{\beta})$ is the probability that a grapheme-color association $(g, c, s) \in \mathbf{X}$ belongs to the subgroup associated with regulatory factor $k \in \mathbf{K}$ (i.e., the probability that the association is due entirely to the influence of regulatory factor k). The conditional probability $\theta_{k,g,c,s}$ is the probability that a grapheme-color association $(g, c, s) \in \mathbf{X}$ that belongs to the subgroup associated with regulatory factor $k \in \mathbf{K}$ has the color $c \in \mathbf{C}$.

Model Estimation

From (3) above, it follows that all values of $\alpha_{k,g,s}(\vec{\beta})$ are completely determined by the values of β and γ. From (2) above, it follows that all values of $\theta_{k,g,c,s}$ are completely determined by the values of π and γ. From (1) above, it follows that all values of γ are completely determined by the values of π. Recall that in this model, all values of π – and thus, all values of γ and θ – are known *a priori* (having been either derived theoretically or determined experimentally).

Therefore, the only unknown parameters in our model are the values of β, and to complete the model requires only that we determine, by maximum likelihood estimation, a best estimate $\hat{\beta}$ for the vector $\vec{\beta}$ of relative RF strengths:

$$\hat{\beta} = \arg\max_{(\vec{\beta})} \mathcal{L}(\vec{\beta} \mid \mathbf{X}) \quad (5)$$

$$\hat{\beta} = \arg\max_{(\vec{\beta})} \prod_{(g,c,s)\in\mathbf{X}} \sum_{k\in\mathbf{K}} \alpha_{k,g,s}(\vec{\beta})\theta_{k,g,c,s}$$

Finally, note that for any vector $\vec{\beta}$ of relative RF strengths and any positive real constant z it follows from (3) above that $\mathcal{L}(\vec{\beta} \mid \mathbf{X}) = \mathcal{L}(z\vec{\beta} \mid \mathbf{X})$. This fact reflects our earlier observation that it is the ratios β_i/β_j of components of $\vec{\beta}$, rather than their individual magnitudes, that are meaningful.

While any one of the infinitely many estimates $z\vec{\beta}$ for $\vec{\beta}$ therefore yields a "correct" model, it is useful (for example, in comparing regulatory factor strength profiles across different datasets, different subsets of regulatory factors, or even different languages) to specify a canonical first regulatory factor K_1 which is always assigned the relative strength $\beta_1 = 1$. For this "reference" or "baseline" regulatory factor we choose the grapheme-independent (and thus, non-linguistic) probability distribution of colors (the "synesthetic palette"; Rouw & Root, 2019), defining corresponding values $\beta_1 = 1, \gamma_{1,g,s} = 1$ for every grapheme $g \in \mathbf{G}$ and subject $s \in \mathbf{S}$, and $\theta_{1,g,c,s} = |\mathbf{X}_c|/|\mathbf{X}|$ for every grapheme $g \in \mathbf{G}$, color $c \in \mathbf{C}$, and subject $s \in \mathbf{S}$, where $|\mathbf{X}_c|$ denotes the cardinality of the set of all associations $x \in \mathbf{X}$ in which some grapheme is associated with the color c. Choosing a specific value for β_1 effectively identifies a unique canonical estimate from the infinite set of equivalent correct estimates $z\vec{\beta}$, resolving any model identifiability issue and completing our model.

Applications of the RF Model

Below, we describe three examples that illustrate how our model can be applied to improve our understanding of how grapheme-color associations are influenced by RFs. We demonstrate that our model can be used to (1) disentangle the effects of RFs that make congruent and incongruent predictions about the same graphemes; (2) directly and quantitatively compare the effect sizes of different RFs using the same statistical analysis; and (3) test for language-dependence of RFs by quantitatively comparing RF strength profiles across synesthetes with different native languages.

Ex. 1: Disentangle Confounded RF Predictions

Univariate analysis of RFs will overestimate effect size when RFs make congruent predictions, and underestimate effect size when RFs make incongruent predictions. For example, the prelinguistic "Basic Shape" RF uses shape-color associations in infants to predict Z to be black (Spector and Maurer, 2011), and the "Index Route" RF *also* predicts Z to be black via semantic associations ("Z is for zebra and zebras are black or white"; Mankin & Simner, 2017). When each RF is modelled separately, each black Z will be "double counted", potentially inflating estimates of effect size. Here, we simultaneously model the influence of the Basic Shape, Index Route, and Color Term RFs to associations on synesthetic associations. We chose these RFs (rather than letter frequency, refrigerator magnets, etc.) not because they are uniquely suited for our model, but because they are straightforward to explain and to operationalize. Our model can include *any* RF that can be specified as a set of probability distributions of predicted color for each grapheme.

Dataset We use the database of synesthete and non-synesthete associations in English and Dutch speakers from Rouw and Root (2019), which is publicly available (https://doi.org/10.6084/m9.figshare.9830816.v1). For the following analyses, only data from English-speaking synesthetes is analyzed. The data are from 54 synesthetes $\mathbf{S} = \{\boldsymbol{s_1}, \boldsymbol{s_2}, \boldsymbol{s_3}, \dots, \boldsymbol{s_{54}}\}$, who were recruited in university classrooms and using posted advertisements at the University of California San Diego (see Rouw and Root, 2019 for further details). The set of synesthetic grapheme-color associations $\mathbf{X}$ for subjects $\boldsymbol{s} \in \mathbf{S}$ and the English graphemes $\mathbf{G} = \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \dots, \mathrm{Z}\}$ was collected using the Eagleman Synesthesia Battery (Eagleman et al., 2007), and synesthesia status was confirmed using the test-retest consistency threshold in Rothen et al. (2013). RGB color data from the Eagleman Battery were transformed into a set of colors $\mathbf{C}$ by using a lookup table (Jraissati and Douven, 2018) to convert RGB values into the 11 Basic Color Terms (Berlin and Kay, 1991):

$$\mathbf{C} = \begin{Bmatrix} black, white, red, green, yellow, blue, \\ orange, brown, purple, pink, grey \end{Bmatrix}$$

Regulatory Factors The Basic Shape RF was operationalized as follows: for all subjects $\boldsymbol{s} \in \mathbf{S}$,

$$\pi_{k=\text{"Basic Color"},g,c,s} = \begin{cases} 1 \text{ if } g \in \{I, O\} \text{ and } c = white \\ 1 \text{ if } g \in \{X, Z\} \text{ and } c = black \\ 0 \text{ otherwise} \end{cases}$$

Note that the Basic Shape RF only makes predictions about a small subset of letters (four). We include it here not only because of its importance to synesthetic associations (it is perhaps the only true language-independent RF; Root et al., *In Review*), but also to illustrate the capability of our model to calculate unbiased RF strength values for RFs that make predictions about subsets of graphemes. Indeed, many RFs in the synesthesia literature only make predictions about a subset of graphemes (e.g., vowels), so is a critical feature of any model that seeks to account for RFs. This is also the feature of our model that distinguishes it from a typical categorical mixture model: in the calculation of $\alpha_{k,g,s}(\vec{\beta})$ (Eq. 4), $\gamma_{k,g,s}$ is zero when an RF makes no prediction about a grapheme, and thus $\hat{\beta}$ estimates the strength of each RF based only on the graphemes for which it makes a prediction.

The Index Route RF was operationalized following the methods of Mankin and Simner (2017), using the data from Root et al. (*In Review*). 65 non-synesthetes typed the first five words that "came to mind" for each grapheme $g \in \mathbf{G}$. From this data, we calculated the frequency $P(w|g)$ of the three most frequently chosen words for each grapheme $g \in \mathbf{G}$, which we call the "index words" for that grapheme. Next, 47 non-synesthetes (new subjects) chose the "best" Berlin-Kay

color $c \in \mathbf{C}$ for each index word. From these data, we calculated the probability $P(c|w)$ that subjects associated each color with each index word. From these data, for all subjects $s \in S$, $\pi_{k=\text{"Index Route"},g,c,s} = \sum_{\mathrm{w}} P(w|g)P(c|w)$.

The Color Term RF was operationalized for all synesthetes $s \in \mathbf{S}$ using the ease of generation naming data of Battig and Montague (1969) for English. First, $\theta_{k=\text{"Color Term"},g,c,s}$ was calculated as the conditional probability that Battig and Montague's subjects named the color c when prompted to name a color, given that they named a color that begins with grapheme g. Next, $\gamma_{k=\text{"Color Term"},g,s}$ was calculated as the probability that Battig and Montague's subjects named any color beginning with grapheme g. From this data, it follows that $\pi_{k,g,c,s} = \gamma_{k=\text{"Color Term"},g,s}\theta_{k=\text{"Color Term"},g,c,s}$.

Results Figure 1 depicts the effect size (relative RF strength values $\boldsymbol{\beta}$) for each of the three RFs included in the model.

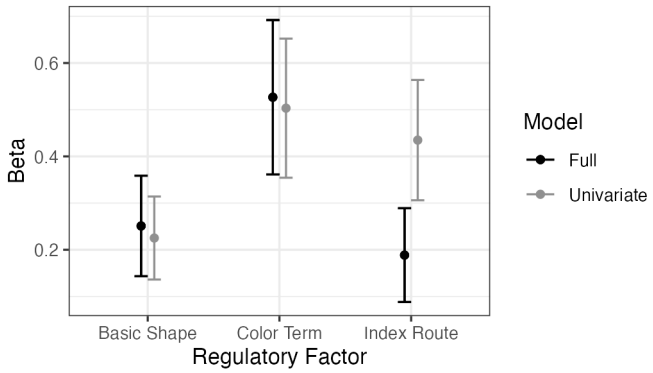

Figure 1: β values for Basic Shape, Color Term, and Index Route RFs, modelled independently (blue) and together (red). Error bars are 95% confidence intervals.

It is clear from these results that the univariate estimate of the Index Route RF effect size is much larger than the effect size estimate of our full model (although the RF is still significantly different from zero). It is likely that previous univariate models of the Index Route underestimated the effect of the overlap in predictions with other RFs (e.g., the Basic Shape and Index Route RFs both predict Z to be black). By accounting for the predictions of all three RFs simultaneously in the same model, we can disentangle confounded RFs and thus obtain uninflated effect size estimates for RFs that make overlapping predictions.

Ex. 2: Compare RF Strength Between RFs

Since most studies of RFs examine only one RF in isolation, the statistical techniques used in the literature vary widely, from simple t-tests and correlations (e.g. Simner et al., 2005) to multiple regression (e.g. Asano & Yokosawa, 2013) to Monte Carlo resampling (e.g. Mankin & Simner, 2017). As a result, it is difficult to compare the relative effect sizes of different RFs as reported in the literature, because the analyses used are so different. For example, Mankin and Simner (2017) report the effect size of the Index Route RF as the number of matches between the two most common synesthetic colors for each grapheme, and the two most common index words for each grapheme. In contrast, Simner et al. (2005) report the effect size of the Color Term RF as the average number of consistent predictions per subject.

By forcing all RFs to be specified in the same way (as a set of predicted probabilities for each grapheme in a shared color space), our model yields effect size estimates that can be directly compared. For example, from Figure 1 and Table 1, the Color Term RF appears stronger than the Basic Shape RF.

Table 1: β values and standard errors

Regulatory Factor	Estimate	Std. Error
Basic Shape	0.25	0.054
Color Term	0.53	0.083
Index Route	0.19	0.050

This observation can be quantified using a two sample Z test of the difference in β coefficients (the standard error of each β can be estimated as the square roots of the diagonal elements of the inverse Hessian matrix). Table 1 lists the β coefficients and standard errors for each RF. We can test for the difference between the effect size of the Color Term and Basic Shape RFs (for example) as follows:

$$z = \frac{0.53 - 0.25}{\sqrt{0.083^2 + 0.054^2}} = 2.83$$

Using this procedure, we find that the Color Term RF is significantly stronger than both the Basic Shape RF ($z = 2.83, p = 0.005$) and the Index Route RF ($z = 3.61$, $p < 0.001$), but the Basic Shape RF is not significantly stronger than the Index Route RF ($z = 0.95, p = 0.34$).

The same general procedure can be used to test if a β coefficient is different from zero – a one sample Z test of the alternate hypothesis that $\beta > 0$. Here, all three β coefficients are significantly larger than zero (Basic Shape RF: $z = 4.66, p < 0.001$; Color Term RF: $z = 6.37, p < 0.001$; Index Route RF: $z = 3.75, p < 0.001$), suggesting that all Regulatory Factors explain some variance in our dataset.

A more rigorous test of significance for βs is a Likelihood Ratio Test (LRT): to test the significance of the coefficient β_k, the likelihood $\mathcal{L}(\vec{\beta} \mid \mathbf{X})$ is compared to the restricted model $\mathcal{L}(\overrightarrow{\beta_0} \mid \mathbf{X})$ in which $\beta_k = 0$. The test statistic $-2\ln(\mathcal{L}(\overrightarrow{\beta_0} \mid \mathbf{X})/\mathcal{L}(\vec{\beta} \mid \mathbf{X}))$ is asymptotically (with sample size) chi-square distributed with degrees of freedom equal to the difference in dimensionality between $\vec{\beta}$ and $\overrightarrow{\beta_0}$ (Wilks, 1938). In this case, using the Likelihood Ratio Test yields identical results to the Wald Test for all three RFs (all $\chi^2 \geq 22.31$; all $p < 0.0001$), confirming that all three RFs explain some variance in our dataset.

By reparameterizing the model in various ways, the LRT can also be used to test other hypotheses. For example, if

there are two alternative operationalizations of a single RF, both operationalizations can be included as separate RFs in the full model, and then two LRTs can be run using two different restricted models (in which the coefficient for each of the two operationalizations is set to zero); if only one operationalization significantly improves the model fit, this suggests it is the more accurate operationalization for the RF.

In sum, our model can be used to compare the relative effect sizes of different RFs in the same dataset. This procedure enables predictions about the relative strength of different RFs to be tested, enables competing hypotheses about the best operationalization for an RF to be compared, and enables all individual RFs to be tested for significance with the same statistical test, thus eliminating confounds that make effect sizes in the current literature difficult to compare.

Ex. 3: Compare RF Strength Between Languages

Many RFs reported in the literature involve linguistic properties (letter frequency, pronunciation, etc.). Despite their clear linguistic origin, these RFs are almost always studied in English-speaking synesthetes, leaving open the question of whether RFs are language-specific or language-universal. Root et al. (2018) provide evidence for at least one potentially-universal RF: across five different languages, synesthetes associate the grapheme in the first ordinal position (e.g., "A" in English) with red. However, some RFs may be language-dependent: Watson et al. (2011) find no evidence that pronunciation influences the associations of English-speaking synesthetes, but Asano and Yokosawa (2013) and Kang et al (2018) find that in Japanese- and Korean-speaking synesthetes, respectively, similarly-pronounced graphemes are associated with similar colors.

Asano and Yokosawa (2013) suggest that differences in orthographic depth between the two languages may drive the Pronunciation RF to be stronger or weaker: English is a deep orthography, whereas the Japanese and Korean orthographies are transparent. However, although this hypothesis is appealing, these three studies are not directly comparable because in each study pronunciation was operationalized in very different ways. This illustrates a third strength of our model: not only can the effect size of different RFs be compared, the effect size of a single RF can be compared across different datasets (for example, of synesthetes with different native languages). Below, we apply our model to a dataset of Dutch-speaking synesthetes, using the same three RFs as in Example 1, and compare the β coefficients for the dataset of Dutch-speaking synesthetes to the β coefficients for English-speaking synesthetes.

Dataset The dataset of English-speaking synesthetes was the same as used in Example 1, above. The dataset of Dutch-speaking synesthetes comes from the same database of synesthete and non-synesthete associations in English and Dutch speakers from Rouw and Root (2019). The data are from 126 synesthetes $\mathbf{S} = \{\mathbf{1}, \mathbf{2}, \mathbf{3}, \ldots, \mathbf{126}\}$, who were recruited from the general public (via television and radio interviews) and from the undergraduate subject pool at the University of Amsterdam (see Rouw and Root, 2019 for further details). The set of grapheme-color associations $\mathbf{X}$ for subjects $\boldsymbol{s} \in \mathbf{S}$ and the Dutch graphemes $\mathbf{G} = \{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \ldots, \boldsymbol{Z}\}$ was collected, verified, and transformed into the Berlin-Kay color space using the same procedure as in Example 1.

Regulatory Factors The Basic Shape RF was identical to that used for English in Experiment 1: as this RF is pre-linguistic, it is identical across languages. The Index Route RF was operationalized in an identical manner to the English version of the RF in Example 1, using data from the same paper (Root et al., *In Review*). 53 Dutch non-synesthetes typed the first five words that "came to mind" for each grapheme $g \in \mathrm{G}$. From this data, we calculated the frequency $\mathrm{P}(w|g)$ of the three most frequently chosen words for each grapheme $g \in \mathrm{G}$, which we call the "index words" for that grapheme. Next, 57 Dutch non-synesthetes (new subjects) chose the "best" Berlin-Kay color $c \in \mathrm{C}$ for each index word. From these data, we calculated the probability $P(c|w)$ that subjects associated each color with each index word. As in Example 1, $\pi_{k=\text{"Index Route"},g,c,s} = \sum_w P(w|g)P(c|w)$.

The Color Term RF was operationalized in an identical manner to the English version of the RF in Example 1. Battig and Montague's (1969) ease of generation experiment was replicated in Dutch by Storms (2001), allowing us to use the exact same procedure as in Example 1 to derive each value of $\pi_{k=\text{"Color Term"},g,c,s}$ using the Dutch color naming data.

Results Figure 2 depicts the effect size (relative RF strength values $\boldsymbol{\beta}$) for each of the three RFs included in the model.

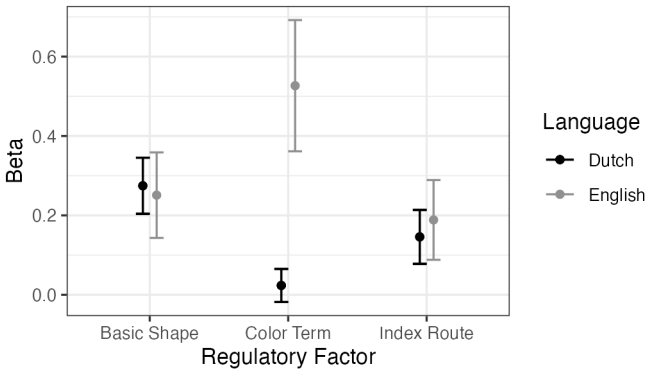

Figure 2: RF strength values for the Basic Shape, Color Term, and Index Route RFs, in English (blue) and Dutch (red) datasets. Error bars are 95% confidence intervals.

The Basic Shape and Index Route RFs do not appear to differ in strength between Dutch and English. We can quantify this observation using the methods from Example 2: the β coefficients did not significantly differ between English and Dutch for the Basic Shape RF ($z = 0.37, p = 0.71$) or the Index Route RF ($z = 0.70, p = 0.48$), but the Color Term RF was significantly stronger in English than in Dutch ($z = 5.90, p < 0.0001$). Furthermore, the strength of the Color Term RF in Dutch was not significantly different from zero (Wald Z test, $z = 1.13, p = 0.26$) and removing the Color

Term RF from the model did not significantly reduce the model likelihood (LRT, $\chi^2(1) = 1.44, p = 0.23$).

Unlike the incongruent result from Example 1, we would not predict *a priori* that the Color Term RF should be present in English and absent in Dutch; indeed, a Color Term RF in Dutch has been described several times (e.g. Rouw et al., 2014; Root et al., 2019). However, previous reports use Simner et al.'s (2007) operationalization (number of matches between color term and color association for each subject), which does not make a prediction about the relative likelihood of two colors that share a first initial. Dutch synesthetes are four times likelier to associate "B" with brown than with blue, whereas our operationalization (using the ease of generation data from Storms, 2001) predicts "B" to be blue much more often than brown.; this discrepancy accounts for our null result here, but does not influence the outcome of Simner et al's (2007) statistical tests. Critically, we believe that a fully specified Color Name RF *should* be able to explain what happens when multiple color terms begin with the same grapheme, thus our null result here reveals that the Color Name RF is not yet fully understood, and alternate operationalizations should be explored.

In sum, our model enables RFs to be fit to data from multiple languages, so that RF strength profiles across language can be compared. Absent other confounds (e.g. poor operationalization), significant differences in RF strength between languages suggest that some property of language might influence the strength of RFs – in other words, that values β may themselves reflect linguistic properties.

Discussion

We created a unifying model of synesthetic Regulatory Factors in grapheme-color synesthesia, that can simultaneously model the contribution of all RFs to grapheme-color associations. Our model controls for confounds in existing univariate models of RFs. For example, RFs that make some congruent predictions lead to inflated univariate estimates of effect size, but by incorporating both RFs into the same model, we can obtain effect size estimates that are not confounded by overlapping predictions. In addition, our model requires all RFs to be specified in the same way, so it is possible to directly and meaningfully compare the effect sizes of each RF: we can make statements like "The Color Term RF is three times stronger than the Index Route RF". Furthermore, our model can be used to compare RF effect sizes across different operationalizations of the same RF, refining our knowledge about which precise properties influence synesthetic associations, and refining our estimates of the relative strength of each RF. Finally, our model can be used to compare RF strength profiles in synesthetes in different datasets, in synesthetes with different native languages, and even in synesthetes vs. control subjects.

It is interesting that RFs explain why synesthetes more frequently experience certain grapheme-color combinations; the fact that many RFs are linguistic in nature suggests that RFs can be used to study language representation in the brain. Indeed, grapheme-color synesthesia should be considered a psycholinguistic phenomenon as much as a perceptual one (Simner, 2007), and the effects of semantic and morphological factors on synesthetic color suggest that the typically modular reading system of the brain is *cognitively penetrable* to synesthesia, and that synesthesia could thus reveal the "specific computations that subserve the reading process" (Blazej & Cohen-Goldberg, 2016). The usefulness of this line of research critically depends on whether linguistic insights to be gained from studying synesthetes (~2% prevalence) generalize to the non-synesthetic population. Indeed, some RFs influence non-synesthetes (Simner et al., 2006), and by comparing RF strength profiles, our model could quantitatively determine exactly which RFs are specific to synesthetes and which RFs influence grapheme-color associations *in general*.

Future work can offer a more standardized method for translating existing Regulatory Factor predictions into the framework of the present model. Some of these translations will involve fitting (or theoretically deriving) additional hyperparameters: for example, to translate the Letter Frequency RF ("more frequent letters are brighter colors") into a set of values of $\pi_{k,g,c,s}$, it is necessary to specify not just the expected brightness for a given frequency, but also the variance in expected brightness; this additional parameter could be fit using a holdout dataset. In many cases, RFs do not make precise predictions about the shape of the predicted probability distribution of colors, and to include these RFs in the model, such precision is necessary. Although this requires incorporating many additional assumptions about RFs, we actually see this as an *advantage* of our model: it forces researchers to make previously implicit assumptions explicit.

Another target of future model development is to decompose the relative RF strength factor β into additional hyperparameters that can be predicted rather than fit. In our model as currently defined, there is no principled prediction of the value of some β_{k}: its fit is determined entirely by maximum likelihood estimation using the dataset of observations **X**. However, in theory, we may be able to predict some amount of variance in each β, and indeed doing so may yield important insights. For example, Asano and Yokosawa (2013) find that pronunciation influences grapheme-color associations in Japanese but not English, and suggest that this difference can be explained by differences in orthographic depth between the two languages (English is a deep orthography, Japanese is transparent). Orthographic depth can be quantified using entropy-based approaches (Borleffs et al., 2017), and the value, in each language, of β_k for the Pronunciation RF might be modelled as a function of grapheme-phoneme entropy. Such a model would be *truly* explanatory in that it would explain not just *how strong* the effect of each RF is, but also *why* the effect is weak or strong.

The framework we develop to model Regulatory Factors in synesthesia might also be useful more generally. In particular, the novel component of our model is the decomposition of the mixture weights α (which can vary across graphemes) into strength parameters β (which do *not* vary across graphemes) and the stimulus specific variable γ,

which is calculated from inputs rather than fit in the model. One advantage of this decomposition is that predictor RFs only contribute to the explained variance for the subsets of graphemes for which they make predictions. This technique could be applicable to model fitting in Pattern Component Modeling (PCM), an extension of the venerable Representational Similarity Analysis (Kriegeskorte et al., 2008) in which mixture weights of feature sets are estimated rather than pre-specified: "[e.g.,] useful if we hypothesize that a region cares about different groups of features (i.e., color, size, orientation), but we do not know how strongly each feature is represented" (Diedrichsen et al., 2018). PCM models take as one of their inputs a set of "feature models", which are analogous to our RFs; formulating the mixture model as specified in the present work would enable feature models to be fit which do not make predictions about neural activity for every stimulus in the dataset. More generally, our framework could be applicable to any variance partitioning scheme (e.g., de Heer et al., 2017) in which an explanatory variable makes predictions for only a subset of the data and does not add noise to the predictions of the remaining data.

Our framework offers a comprehensive, predictive model of synesthetic associations, that yields quantitative estimates of effect size for each synesthetic Regulatory Factor. This brings us closer to one "end goal" of synesthesia research: explaining why a particular synesthete experiences a particular color for a particular grapheme. Furthermore, by linking these predictions to objective, measurable properties of letters, we build an explanatory model of synesthetic associations that offers a window into the multifactorial contributions to letter representation in the brain.

References

Asano, M., & Yokosawa, K. (2013). Grapheme learning and grapheme-color synesthesia: toward a comprehensive model of grapheme-color association. *Frontiers in human neuroscience*, 7, 757.

Asher, J. E., Aitken, M. R., Farooqi, N., Kurmani, S., & Baron-Cohen, S. (2006). Diagnosing and phenotyping visual synaesthesia: a preliminary evaluation of the revised test of genuineness (TOG-R). *Cortex*, 42(2), 137-146.

Battig, W. F., & Montague, W. E. (1969). Category norms of verbal items in 56 categories A replication and extension of the Connecticut category norms. *Journal of experimental Psychology*, 80(3p2), 1.

Beeli, G., Esslen, M., & Jäncke, L. (2007). Frequency correlates in grapheme-color synaesthesia. *Psychological Science*, 18(9), 788-792.

Blazej, L. J., & Cohen-Goldberg, A. M. (2016). Multicolored words: Uncovering the relationship between reading mechanisms and synesthesia. *Cortex*, *75*, 160-179.

Borleffs, E., Maassen, B., Lyytinen, H., & Zwarts, F. (2017). Measuring orthographic transparency and morphological-syllabic complexity in alphabetic orthographies: a narrative review. *Reading and writing*, *30*(8), 1617-1638.

de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The hierarchical cortical organization of human speech processing. *Journal of Neuroscience*, *37*(27), 6539-6557.

Diedrichsen, J., Yokoi, A., & Arbuckle, S. A. (2018). Pattern component modeling: a flexible approach for understanding the representational structure of brain activity patterns. *NeuroImage*, *180*, 119-133.

Kang, M. J., Kim, Y., Shin, J. Y., & Kim, C. Y. (2017). Graphemes sharing phonetic features tend to induce similar synesthetic colors. *Frontiers in psychology*, 8, 337.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 4.

Lovelace, C. T. (2013). Synesthesia in the 21st century: Synesthesia's ascent. In Simner & Hubbard (Eds.), *The Oxford handbook of synesthesia*. Oxford University Press.

Mankin, J., & Simner, J. (2017). A is for apple: the role of letter–word associations in the development of grapheme–colour synaesthesia. *Multisensory research*, 30, 409-446.

Root, N. B., Dobkins, K., Ramachandran, V. S., & Rouw, R. (2019). Echoes from the past: synaesthetic colour associations reflect childhood gender stereotypes. *Philosophical Transactions of the Royal Society B*, *374*(1787), 20180572.

Root, N., Asano, M., Melero, H., …Rouw, R. (*In Review*). Does the color of your letters depend on your language? Language-dependent and universal influences on grapheme-color synesthesia in seven languages.

Rouw, R., Case, L., Gosavi, R., & Ramachandran, V. (2014). Color associations for days and letters across different languages. *Frontiers in psychology*, *5*, 369.

Rouw, R., & Root, N. B. (2019). Distinct colours in the 'synaesthetic colour palette'. *Philosophical Transactions of the Royal Society B*, 374(1787), 20190028.

Simner, J. (2007). Beyond perception: synaesthesia as a psycholinguistic phenomenon. *Trends in cognitive sciences*, *11*(1), 23-29.

Simner, J. (2013). The "rules" of synesthesia. In Simner & Hubbard (Eds.), *The Oxford handbook of synesthesia*. Oxford University Press.

Simner, J., Ward, J., Lanz, M., … Oakley, D. A. (2005). Non-random associations of graphemes to colours in synaesthetic and non-synaesthetic populations. *Cognitive neuropsychology*, 22(8), 1069-1085.

Spector, F., & Maurer, D. (2011). The colors of the alphabet: Naturally-biased associations between shape and color. *Journal of Experimental Psychology: Human Perception and Performance*, 37(2), 484.

Storms, G. (2001). Flemish category norms for exemplars of 39 categories: A replication of the Battig and Montague (1969) category norms: *Brain*, *124*, 1619-1634.