Quantifying Visual Preferences Around the World

Katharina Reinecke
University of Michigan
Ann Arbor, MI 48109
reinecke@umich.edu

Krzysztof Z. Gajos
Harvard University
33 Oxford St., Cambridge, MA
kgajos@eecs.harvard.edu

ABSTRACT
Website aesthetics have been recognized as an influential moderator of people's behavior and perception. However, what users perceive as "good design" is subject to individual preferences, questioning the feasibility of universal design guidelines. To better understand how people's visual preferences differ, we collected 2.4 million ratings of the visual appeal of websites from nearly 40 thousand participants of diverse backgrounds. We address several gaps in the knowledge about design preferences of previously understudied groups. Among other findings, our results show that the level of colorfulness and visual complexity at which visual appeal is highest strongly varies: Females, for example, liked colorful websites more than males. A high education level generally lowers this preference for colorfulness. Russians preferred a lower visual complexity, and Macedonians liked highly colorful designs more than any other country in our dataset. We contribute a computational model and estimates of peak appeal that can be used to support rapid evaluations of website design prototypes for specific target groups.

Author Keywords
Website Aesthetics; Colorfulness; Complexity; Modeling; Adaptation; Personalization

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): User Interfaces

INTRODUCTION
While the field of human-computer interaction has traditionally been mostly concerned with functionality and usability, aesthetics are increasingly regarded as an additional dimension that "augments other aspects of the design and the overall interactive experience" [27, p. 4]. Aesthetics have been recognized as important because of their positive influence on people's behavior, such as on performance under conditions of poor usability [18], or on purchase intentions [4]. Even before elaborate considerations about purchases can possibly take place, the first impression of appeal determines how we perceive other attributes of a product, such as its usability and trustworthiness [15, 14].

With these potentially long-lasting consequences of people's first impressions in mind, it would be desirable to specify what constitutes "good design". However, it has been argued that universal design guidelines are useful only to a certain extent, because aesthetic impressions vary substantially across individuals [16, 15]. What someone finds appealing seems to depend on individual and demographic differences, such as personality, gender, or age [20, 6, 24, 12]. To maximize website appeal for a given user, it would therefore be best to offer designs personalized to their individual visual preferences.

In this paper, we address one of the main challenges for achieving such personalized website designs: Knowing what a user likes. Our goal is to establish a better understanding of how people's individual visual preferences differ and how we can best predict them.

To achieve this goal, we conducted an online study on our experimental platform LabintheWild.org and collected approximately 2.4 million subjective ratings of visual appeal from almost 40 thousand participants of diverse backgrounds applied to a set of 430 websites. Building on the work of [25], we used a set of computational image metrics and perceptual models to estimate each website's colorfulness and visual complexity. We then used the collected subjective ratings to characterize how colorfulness and visual complexity impact subjective perceptions of visual appeal based on age, gender, geography, and education. For each subgroup, we additionally calculated the most highly preferred levels of colorfulness and visual complexity. The analyses of the differences in the estimates of peak appeal demonstrate that there are substantial differences in people's first impressions of aesthetics, and that geographic location, age, gender, and education level all play a significant role in determining their preferences. Measuring the differences between these peaks of appeal for various subgroups, we found, for example, that females like websites with highly saturated colors more than males. Education level negatively correlates with preferences for colorful and complex sites. Finland and Russia are among the countries whose members are most negatively affected by a high visual complexity, and Macedonians prefer the most colorful websites of all countries in our dataset.

We make three main contributions:

(1) We identify several demographic factors that impact people's visual preferences, and characterize how they influence appeal by pointing out several between-group differences in

CHI'14, April 26–May 1, 2014, Toronto, Canada.

http://dx.doi.org/10.1145/2556288.2557052

visual preferences. This analysis is the first to report on a largely heterogeneous sample in terms of age, occupation, education, geographic location, and web experience, contributing new knowledge about previously understudied groups.

(2) We developed a computational model that combines a user's demographic information with computational image metrics for assessing a website's colorfulness and complexity in order to predict a user's subjective perception of visual appeal. Our model improves upon a recently proposed universal model [25], which did not enable predictions for specific demographic groups due to a relatively small sample (242 participants). With this larger sample, we provide estimates of the level of colorfulness and visual complexity at which appeal peaks for specific demographic groups. The model and peak estimates can support rapid evaluations of the suitability of website design prototypes for certain target groups.

(3) We contribute the first public dataset on visual appeal including the preference ratings of almost 40,000 participants for 430 website screenshots and self-reported demographic background information.

In the following two sections, we describe related work on aesthetics and the computational image metrics on which we base our analyses. The second half of the paper then describes our data collection method, experiment design, and analyses. We present our model results, highlight several between-group differences in people's visual preferences, and point out website designs that were found to appeal to certain subgroups. The paper concludes with a discussion, future research, and directions for how to access the dataset.

RELATED WORK

Most previous aesthetics-related research in HCI has focused on finding universal design guidelines. In an effort to objectively measure aesthetics, Zheng et al. [31] excluded participants' ratings for website screenshots seen for 150ms if they were more than 2 standard deviations away from the mean ratings of their 22 participants. Other work attempting to quantify aesthetics assumed that websites that received a Webby Award (selected by expert judges) must constitute good website design [13].

Contrasting the idea of universal designs, research has repeatedly found large individual differences in aesthetic judgements [16, 15, 14]. These variabilities have been partially attributed to people's sensitivity for aesthetics [11], and to differences in demographic backgrounds [28]. In a first attempt to automatically predict users' aesthetic preferences, Reinecke et al. [25] introduced a model of website aesthetics based on a number of image metrics. While they found that several demographic variables impacted appeal, a relatively small sample size of 242 participants prevented more in-depth analyses of differences in people's aesthetic taste. Difficulties in recruiting larger and more diverse samples might also be the reason why most other work in this direction has focused on only one demographic difference at a time. Tuch et al. [29], for example, showed that males and females differ in their aesthetic reaction towards asymmetrical websites while reporting on a relatively homogeneous sample of 60 undergraduate psychology students in Switzerland. In a study with UK university students, Moss and Gunn [20] further found that female participants preferred websites designed by females over those designed by males, and vice versa for males. In addition, Hsiu-Feng [12] observed gender differences in Taiwanese children between 12 and 14 years of age. Girls preferred a low to medium visual complexity in websites, whereas boys preferred medium to highly complex sites. Investigating the influence of education level, Chen et al. [5] found that website aesthetics differ between students in their first and those in their final year. Furthermore, several researchers have investigated whether cultural background determines preferences (see, e.g., [1, 7, 6, 24]).

These previous studies suggest that gender, geographical location, and education level might play a role in people's aesthetic preferences. However, none of the studies on website aesthetics have compared several age groups, a variety of different education levels, the impact of Internet usage, a large number of countries, or whether differences in preferences between genders also hold across different age groups. An additional shortcoming of previous studies is that they are difficult to compare: First, their relatively small sample sizes inhibit broader findings, such as how demographic factors interact with each other. In addition, low numbers of stimuli hinder generalizability, or reproducibility of findings with other populations and/or stimuli. This work therefore extends prior work with an analysis of a larger and more diverse sample and set of website stimuli. We also contribute a computational model of website aesthetics that is based on a set of image metrics described next.

QUANTIFYING APPEAL

While a method to quantify website appeal in all its facets has yet to be developed, researchers have focused on two of the most prominent website characteristics: colorfulness and visual complexity. The choice of colors has been shown to affect a website's perceived trustworthiness and users' loyalty [18, 6, 14]. The composition, number, and choice of colors (i.e., the overall colorfulness of a website) influences appeal [25]. Visual complexity, however, is often thought to be the greater predictor of appeal [17, 31, 30]. Sometimes described with the negatively connoted term "clutter" [26], recent research shows it does indeed negatively relate to appeal [30]. In contrast, Berlyne's influential theory on visual complexity [3] suggests an inverted U-shape relationship, where moderately complex stimuli are found most appealing.

By computing low-level image statistics for website screenshots, Zheng and colleagues [31] demonstrated that it is possible to approximate participants' perception of complexity. Reinecke and colleagues [25] extended Zheng et al's set of image metrics and evaluated how this larger set predicts people's perceived colorfulness and complexity of websites. With ratings from several hundreds of participants, they developed two computational models of perceived colorfulness and complexity of websites.

Here we employ Reinecke et al.'s perceptual models to assess each website's colorfulness and complexity. Their colorfulness model was based on a computation of the average saturation of colors across all pixels of a website screenshot,

a calculation of colorfulness following [10], the number of image areas, the number of leaves resulting from a quadtree decomposition algorithm (enabling an analyses of the spatial distribution of colors), the number of areas identified to contain text or other objects as per a space-based decomposition of the webpage, as well as the percentage of pixels containing one of seven colors. The visual complexity model included the number of areas containing text or other objects, the number of leaves resulting from a space-based decomposition, the number of text groups (e.g., a paragraph), the number of images, a computation of colorfulness based on the average saturation, and hue.

The procedure to compute these image metrics is described in [31] and [25]. The set of algorithms takes a 1024x768 sized website screenshot as input and outputs the values for each image metric. As a sanity test, after applying these metrics to our 430 website stimuli, we ranked a random selection of website screenshots according to each of these values separately in order to visually evaluate the correctness of the image metrics and the two perceptual models. While the model-generated ordering of websites according to the colorfulness and complexity models was reasonable and perceptually intuitive, the values computed for symmetry and balance (following Zheng et al.'s description in [31]) did not correspond to our perceptual judgments. These spatial metrics were therefore left out in our analysis.

EXPERIMENT

We designed this experiment with two main goals in mind: First, to compare participants' first impressions of website aesthetics across a variety of demographic backgrounds, and second, to develop predictive models that would account for demographic differences in the perception of visual appeal.

Method

Materials

Our stimuli consisted of a set of 430 website screenshots divided into 350 English language websites, 60 foreign websites (using a different writing system), and 20 websites that had been nominated for the Webby Awards in recent years. Websites were selected to not have received wide public exposure, to represent a large variety of genres, and to include a range of colorfulness and visual complexity levels.

Procedure

The study was designed as a 10-minute online test and launched on our experimental platform LabintheWild.org to achieve the diversity (in terms of geographic location, age, education, socio-economic status, and web experience) that is needed to study aesthetic preferences for websites across diverse demographic groups. Participants did not receive monetary compensation, but were instead incentivized with a comparison of their visual preference results to others. Following the experimental procedure in [14] and [25], participants were asked to rate screenshots of websites on perceived visual appeal on a scale from 1 to 9. Screenshots were displayed for 500ms to capture participants' first impression of the websites' aesthetics and minimize the influence of their content. All stimuli were downsized from their original 1024x768 screenshot size to 600 pixels in width, and presented on a white background.

After giving their informed consent, participants were asked to fill out a demographics questionnaire. They then received instructions about the experiment, and were able to test this by rating a fixed practice set of five website screenshots (shown in random order). We used the same five websites to anchor all participants' ratings. The ratings from this test phase were not included in the analysis. As a next step, participants rated a stratified random sample of 30 websites (22 in English language, 4 foreign, and 4 Webby Award websites) presented in random order and drawn from the larger pool of 430 websites. The second evaluation phase presented participants with the same 30 websites (again presented in random order) to control for consistency in participants' ratings. Instructions were presented in English.

Participants

We report on data collected between June 2012 and August 2013. During this time, 39,975 volunteers (54% female) from 179 countries completed the experiment on LabintheWild.org. Forty percent had lived in another country for at least 6 months, and/or had parents of a different nationality. Participants were between 12 and 91 years old (mean = 32.4, sd=12.8 years) and 41% had at least some college education. We additionally collected information about countries of residency in the order from birth to present, the duration spent in each country, native and learned languages, as well as fluency levels, current residency in an urban, suburban, or rural setting, education level, web usage (number of hours per day), and profession.

Data Preparation and Analyses

Participants who reported that they did not have normal or corrected-to-normal vision or that they had previously participated in the study were excluded from the analysis. We also omitted participants who did not fill in the demographics, or who reported countries for their own or their parents backgrounds that suggested random picking from the top of the list (e.g., Antarctica, or combinations such as Angola, Antilles, Aruba), as well as participants whose sum of years spent in different countries was hugely different from their age. Finally, we omitted the data of participants under 12 years of age and over 91, where our sample size significantly dropped.

We then analyzed the consistency in participants' ratings across phase 1 and 2 of the experiment, and omitted 47,510 rating pairs (4.9% of all observations) that differed by more than 2 points on the 9-point Likert scale. The resulting standard deviation of the difference between participants' ratings in phase 1 and 2 is 0.7, indicating that their ratings are reliable and representative of their preferences. The cleaned data includes 1,542,166 observations (771,083 paired ratings from 32,222 participants).

In the analyses of country influences, we additionally omitted culturally ambiguous participants who had lived in multiple countries in their lives (at least six months in another country) or whose parents were of a different nationality. We further excluded all countries with fewer than 1000 paired ratings.

For these analyses, the dataset consists of 441,478 paired ratings and 43 countries (from 18,448 participants). The majority of participants in this data set were from the US (43%), followed by the UK (17%), Hungary (6%), Canada (5%), and Romania (3%, all others $\leqslant 3\%$) When referring to average visual appeal scores, we therefore report the marginal means after having controlled for gender, education, country, and age to account for differences in the distribution of demographics.

To analyze the impact of demographic variables on aesthetic preferences for websites, we fitted a series of linear mixed-effects models representing the hierarchical structure of our data. All models were constructed using *R* and the package *lme*4 [2]. First, a null model (intercept/empty) containing only Website ID and Participant ID as random factors was fitted. We then fit a model adding colorfulness and visual complexity as fixed factors. Both of these website characteristics were computed based on the models presented in [25], with scores ranging from 1 to 9 for colorfulness and 1 to 10 for visual complexity. Colorfulness and visual complexity were included with their linear and quadratic terms. Finally, we fitted a full model, which additionally included demographic variables (selected if previous literature provided a basis for inclusion). Demographic variables were modeled as interaction terms with both colorfulness and visual complexity.

After this initial model construction, we repeated the regression excluding variables that were statistically non-significant. Demographic variables were included or omitted based on Akaike's Information Criteria (AIC), which provides an estimate of the relative fit of alternative models. According to this procedure, age, gender, country, and education significantly improved the model fit; all other demographic variables were excluded from the model.

To quantify the absolute model fit and obtain information about the variance explained by the model, we calculated the marginal R^2 (the variance explained by fixed factors), as well as the conditional R^2 (the variance explained by fixed and random factors combined) following a newly developed procedure [21]. The difference between conditional and marginal R^2s explain how much variability is in the random effects Website ID and Participant ID. When referring to the regression results, we report on the F statistics from the Analysis of Variance table. The full list of model parameters and regression coefficients, as well as detailed information on how variables were coded and entered into the regression can be accessed at http://iis.seas.harvard.edu/resources/.

Finally, to estimate the mean values and standard errors of peak appeal (e.g., the complexity or colorfulness levels at which appeal ratings were the highest), we used a bootstrap procedure that has previously been identified as a valid approach to compute peak estimates [9]. The procedure randomly resampled the data 1000 times (with replacement). The resampling was done on a per-website basis. For each bootstrap iteration, we fit a Lowess (locally weighted scatterplot smoothing) curve with the smoothing parameter $\alpha = .25$. For each bootstrap sample we computed the complexity/colorfulness score for which the Lowess function peaked. The mean of these estimates is the bootstrap estimate of the optimal colorfulness/complexity (the peak appeal, cf. Figure 1). The standard deviation of those samples is the estimate of the standard error of the estimate of the mean.

We additionally calculated Cohen's d as a measure of effect size for the comparisons between pairs of means, such as to compute the standardized mean difference between mean appeal at low and high complexity, or the difference between mean appeal at low complexity and the peak, as exemplified in Figure 1. Low/high complexity/colorfulness websites are the 5% of sites with the lowest/highest complexity/colorfulness scores. Peak websites are 0.5 complexity/colorfulness scores lower/higher than the score of peak appeal. While we intermittently refer to these values where needed, the complete list of Cohen's d values can be accessed at http://iis.seas.harvard.edu/resources/.

General Results

A comparison of our null model with the full model using the likelihood ratio test showed that the full model fits the data significantly better ($\chi^2_{(294)} = 7510.9, p < .0001$). In addition, comparing the model without demographics (using only colorfulness and visual complexity as fixed effects and participant ID and website ID as random factors) to our full model showed that age, gender, geographic location, and education level significantly improve the model fit ($\chi^2_{(290)} = 7408.7$, $p < .0001$). In other words, demographic variables play a significant role in supporting the prediction of visual appeal. Visual inspection of residual plots showed that the data meets the assumptions of homoscedasticity and normality.

Our final model explains 47% of the variance in people's first impressions of appeal based on a website's visual complexity and colorfulness (conditional $R^2 = .47$, marginal $R^2 = .07$, see also http://iis.seas.harvard.edu/resources/ for details on the model). The result is comparable to that of Reinecke et al. [25] despite the fact that we worked with a much more heterogeneous sample.[1]

In the following, we first report on preferences for visual complexity and colorfulness in general before showing how these results are influenced by demographic factors.

General Results for Visual Complexity

Our results show that the perceived visual complexity of websites is a strong predictor of appeal ($F_{(1)} = 23.96$, $p < .001$), which confirms previous findings about the role of visual complexity on users' first impressions (e.g., [19, 30]). In line with the results of [25], visual complexity plays a more important role as a predictor of appeal than colorfulness.

As suggested by Berlyne [3] and consistent with some of the previous studies (e.g., [8, 25]), the relationship between visual appeal and complexity in our data is best described by

[1] Note that our R^2 cannot be directly compared to Reinecke et al.'s adj. R^2 of .48 [25], because unlike most previous calculations that use the maximum likelihood estimates of the model parameters, Nakagawa et al.'s new method for calculating R^2 [21] does not disregard the uncertainties around these estimates.

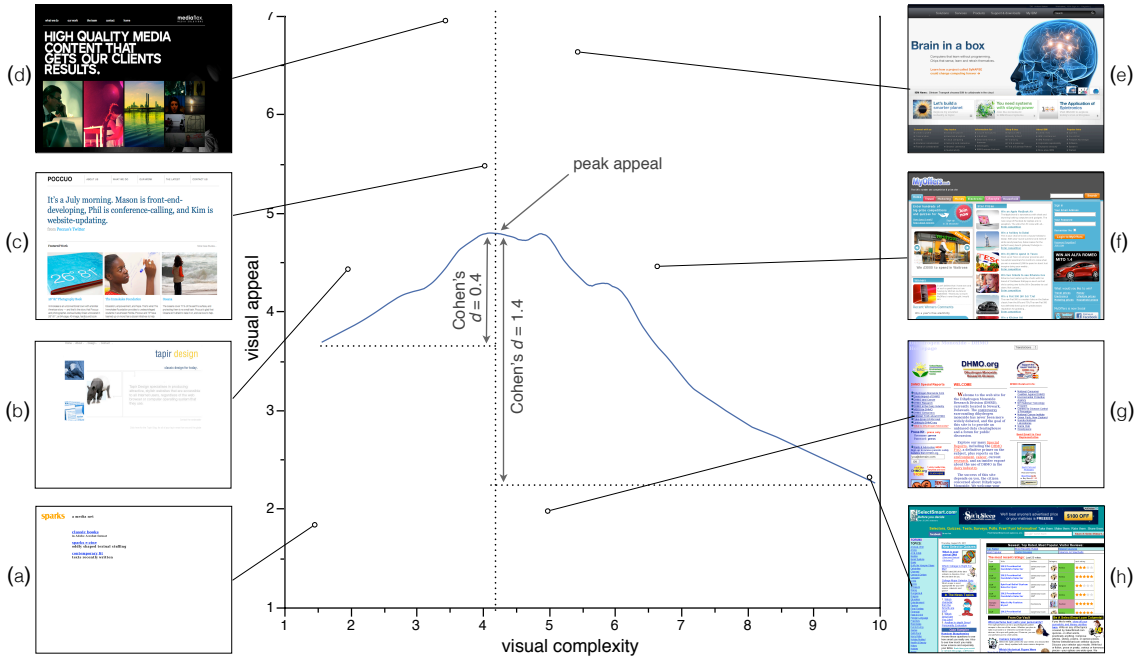


Figure 1. The relationship between visual complexity and appeal shown with the Lowess curve, peak appeal, and effect sizes calculated with Cohen's d. Website images point to their marginal mean rating of visual appeal after controlling for country, age, gender, and education (standard errors < 0.06) and their visual complexity score as calculated using the perceptual model developed in [25].

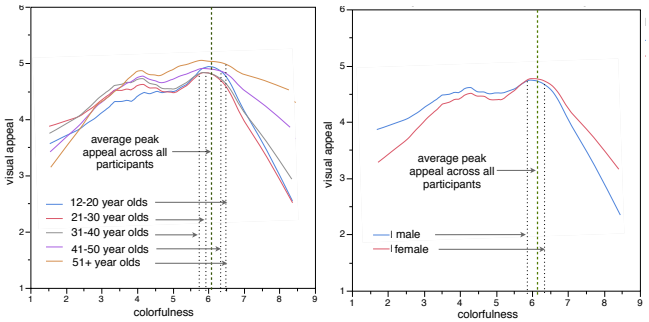
an inverted U-shape (see Figure 1). For our participant sample, visual appeal peaks at a complexity level of 4.2 (out of 10), suggesting that the average person in our dataset prefers websites with a moderate level of complexity. We will later see that the peaks vary between 2.5 and 4.8 depending on demographic background.

The effect sizes (visualized in Figure 1) demonstrate that a low visual complexity has less severe consequences than high complexity does: There was a moderate difference between the average appeal of websites with optimal complexity scores (4.2 ± 0.5) and appeal ratings for websites with low complexity levels (Cohen's $d = 0.6$), but a large difference between peak appeal and the appeal ratings of high complexity levels (Cohen's $d = 2.0$). In other words, appeal steadily declines after a complexity level of around 4.2. This is later than suggested by the negative linear relationship between visual complexity and appeal found by [17] and [30], whose results did not support a quadratic relationship. The discrepancy is almost certainly due to our larger dataset, which included less complex stimuli (see, e.g., Figure 1(a)) than those those that Tuch et al. [30] used as examples for low complexity. Thus, we conclude that websites with a low to medium complexity, but not extremely low, will appeal to most.

Figures 1(c), (d), and (e) demonstrate examples of websites with an optimal level of complexity and high ratings of appeal. The website in Figure 1(g) has a similar complexity level, but received much lower ratings on appeal. This suggests that the complexity score alone cannot fully predict appeal. A visual analysis of our websites suggests that a moderate complexity results in highest appeal ratings when achieved with a good balance between text, color, and images.

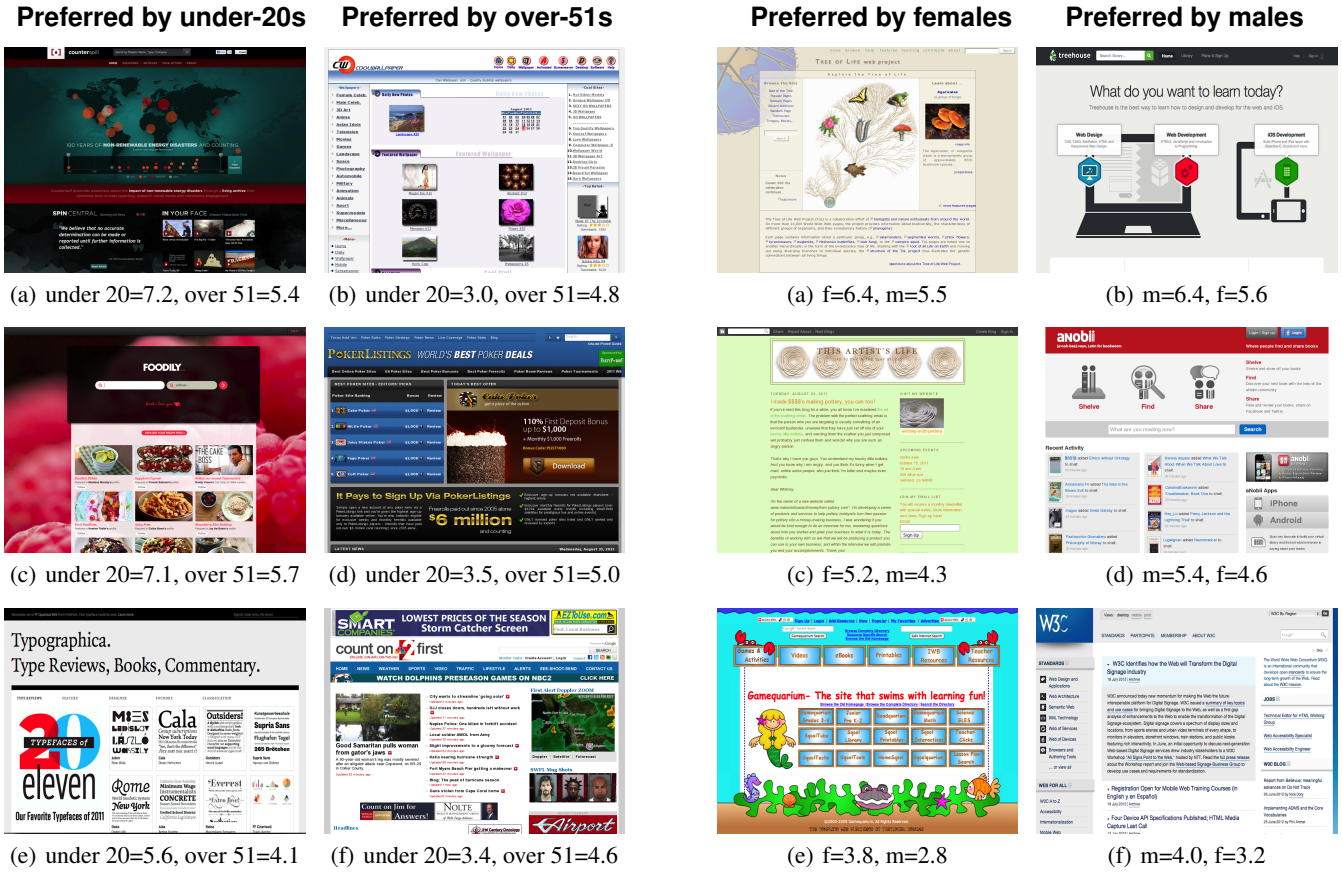
General Results for Colorfulness

We found a significant main effect of colorfulness on appeal ($F_{(1)} = 6.18$, $p < .001$). As demonstrated in Figure 2, the relationship between colorfulness and appeal can be approximated by a similarly inverted U-shape as it was the case for complexity. The average participants' visual appeal for colorfulness peaks at a colorfulness of 6.1 (of 9), but the difference in mean appeal between low/high colorful websites and the peak of appeal is lower than the differences that we saw for complexity (Cohen's $d = 0.3$ and $d = 0.8$). In contrast to complexity, people do not seem to respond as dramatically to slight changes in the level of colorfulness; the lower variability in the ratings of appeal between different levels of colorfulness also means that colorfulness is an overall less informative predictor for appeal than complexity.



(a) Age differences (b) Gender differences

Figure 2. The relationship between colorfulness and appeal for different age groups and genders shown with the Lowess curve. Colorfulness is calculated using the perceptual models developed in [25].



(a) under 20=7.2, over 51=5.4 (b) under 20=3.0, over 51=4.8 (a) f=6.4, m=5.5 (b) m=6.4, f=5.6

(c) under 20=7.1, over 51=5.7 (d) under 20=3.5, over 51=5.0 (c) f=5.2, m=4.3 (d) m=5.4, f=4.6

(e) under 20=5.6, over 51=4.1 (f) under 20=3.4, over 51=4.6 (e) f=3.8, m=2.8 (f) m=4.0, f=3.2

Figure 3. Examples websites with some of the largest differences in mean ratings of appeal between two age groups. Websites preferred by under 20 year olds on the left and those preferred by participants over 50 years of age on the right. All standard errors $\leq$ 0.05.

Figure 4. Example websites with some of the largest differences in mean ratings of appeal between genders. Websites preferred by females more than by males on the left. All standard errors $\leq$ 0.05, f=mean ratings by females, m=mean ratings by males.

According to our results, website designs that appeal to most have a medium to high colorfulness, but a low to medium visual complexity. A good example for this is the website shown in Figure 1(d) with it's complexity level of 3.3, and higher colorfulness of 5.7. Seeing that saturation has a significant influence on the overall perceived colorfulness [25], the finding reaffirms that of Palmer and Schloss [23], who found that (Western) adults prefer colors of higher saturation.

Results on the Influence of Demographics

Our model suggests that preferences are simultaneously influenced by multiple aspects of our demographic backgrounds. In the following, we will attempt to disentangle these effects and point out specific trends within demographic subgroups in the order of importance they play in the model.

Age

Colorfulness significantly interacts with age ($F_{(1)} = 198.3$, $p < .001$). Calculating the peak appeal per age group, we found that participants aged 31 to 40 years prefer a slightly lower colorfulness than others (peak appeal = 5.6, $SE = 0.04$, see also Figure 2(a)). Participants under 20 and those over 51 years of age gave highest ratings for websites with a colorfulness level of 6.5 ($SE < 0.03$). While these peak preferences for a medium to high colorfulness level only slightly differ between age groups (all means of peak appeal between 5.6 and 6.5), older participants find plain, colorless websites less visually appealing than any other age group (Cohen's d between low colorfulness and peak appeal = 1.8 vs. 0.4–1.1 for other age groups), and are less negatively effected by a high colorfulness (cf. Figure 2(a)).

Participants' age also significantly affected their preference for certain levels of visual complexity ($F_{(1)} = 1721.1$, $p < .001$): The older someone is, the more complex they prefer websites to be. The difference in appeal between optimal and suboptimal complexity levels within age groups is large, suggesting that people are more negatively affected by suboptimal complexity levels than by suboptimal colorfulness levels. Participants between 12 and 40 do not strongly differ in their preference for a moderate complexity (peaks for the three different age groups between 4.1 and 4.2, $SE = 0.1$). However, as participants get older, the peak appeal occurs at an increasingly higher visual complexity: For the 41-50 year olds at 4.5 ($SE = 0.1$), and for the over 50 year olds at 4.7 ($SE = 0.01$). In other words, participants over 41 liked websites with a higher complexity than under 40 year olds. This is different from the results of [25] who reported that participants older than 45 years preferred a low visual complexity more than other age groups. We attribute the difference to

our more heterogeneous, and, in particular, on average less educated sample.

Figure 3 visualizes the differences between age groups and their preferences with example websites from our dataset. Websites on the left were rated significantly higher by the youngest age group, and websites on the right significantly higher by the oldest age group (differences in preference are all statistically significant at $p < .001$ when analyzed with independent-sample t-tests and using marginal means controlling for country, education level, and gender). An obvious difference is that websites preferred by under 20 year olds (on the left) make use of saturated colors and larger images. Those websites preferred by older participants are more text-heavy and complex, but use less saturated colors. The complexity model captures this difference by basing its scores on the number of areas containing text or other objects, and on the use of colors.

Gender

The effect of colorfulness on appeal is also moderated by gender ($F_{(1)} = 658.9$, $p < .001$), such that females rate colorful websites higher on visual appeal than males. Figure 2(b) shows the different colorfulness preferences between females and males. Appeal was estimated to peak at a colorfulness value of 5.8 for males ($SE = 0.04$) and at 6.3 for females ($SE = 0.03$).

We also found a significant interaction between visual complexity and gender ($F_{(1)} = 17.8$, $p < .001$), albeit not as strong as for colorfulness. In fact, visual appeal peaks at a similar complexity level of 4.2 for males ($SE = 0.01$) and 4.3 for females ($SE = 0.01$). Where the variation is most pronounced is in the acceptance of websites with a low visual complexity, which females dislike more than males.

Figure 4 shows how this difference between males' and females' preferences manifests itself in some of the most controversial websites between genders. With a high colorfulness score of 8.4 and a complexity score of 4.4, the website in Figure 4(e) is mostly disliked by our participants independent of their gender. Females, however, seem to be more accepting of this design as shown by their average rating of 3.8 (versus 2.8 by males). In contrast, Figure 4(f) shows a website with lower than ideal levels of colorfulness (score of 3.7). The average male likes such simpler look more than females (marginal mean ratings: males = 4.0, females = 3.2).

Anecdotally, what appears to differentiate male-preferred websites from those that females prefer is that male-preferred sites predominantly use saturated primary colors on gray or white background to contrast different regions or items on the webpage. This can be seen in Figure 4(b), 4(d), and (less so) 4(f), and is consistent with other male-preferred websites in our dataset. Websites preferred by females use more homogeneous color schemes and rarely employ stark contrasts between colors to structure content and differentiate elements (see Figure 4(a) and 4(c)).

What we did not expect is that some websites that are meant to appeal to both genders use such "gender-biased" designs and thereby lower the appeal for one of the genders. The website pictured in Figure 4(a), for example, introduces the Tree of Life Web Project, which provides interested biologists with information about biodiversity. Clearly, this site does not primarily target females, although its design appeals much more to women than to men. Similarly, Figure 4(d) shows a website that enables people to "shelve" their books, share book reviews, and find new books. Again, this site most likely intends to target both genders equally, but the design primarily appeals to men.

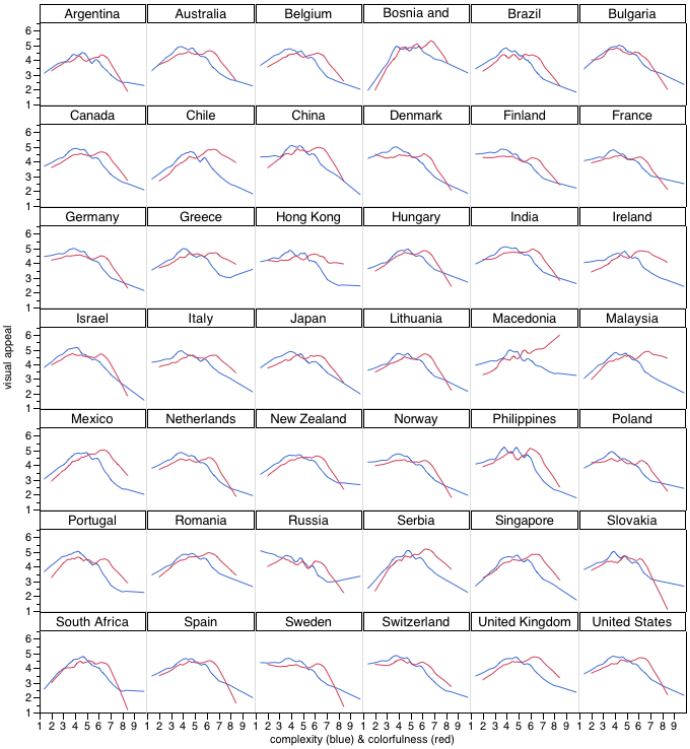


Figure 5. Colorfulness and complexity preferences in example countries. The red and blue lines represent the Lowess curves for the perceived colorfulness and perceived visual complexity, respectively. Both are calculated using the perceptual models developed in [25].

Country

The following analyses report on a subset of our data with 43 countries for which we had collected at least 1000 paired ratings.

The results show a significant interaction between a website's visual complexity and country ($F_{(32)} = 20.50$, $p < .001$). Figure 5 provides an overview of the colorfulness (red) and visual complexity (blue) Lowess curves for 42 example countries (Austria has been omitted for space-saving reasons). While preferences for certain levels of visual complexity show similar U-shape relationships for each country, Figure 6 demonstrates considerable shifts in the peak of appeal. Most notably, participants from Russia preferred the lowest visual complexity of all countries (peak appeal = 2.5, $SE = 0.04$). In contrast, participants from Serbia, Bosnia and Herzegovina, Mexico, and Chile most preferred websites with substantially higher complexity scores between 4.6 and 4.8 ($SE < 0.03$).

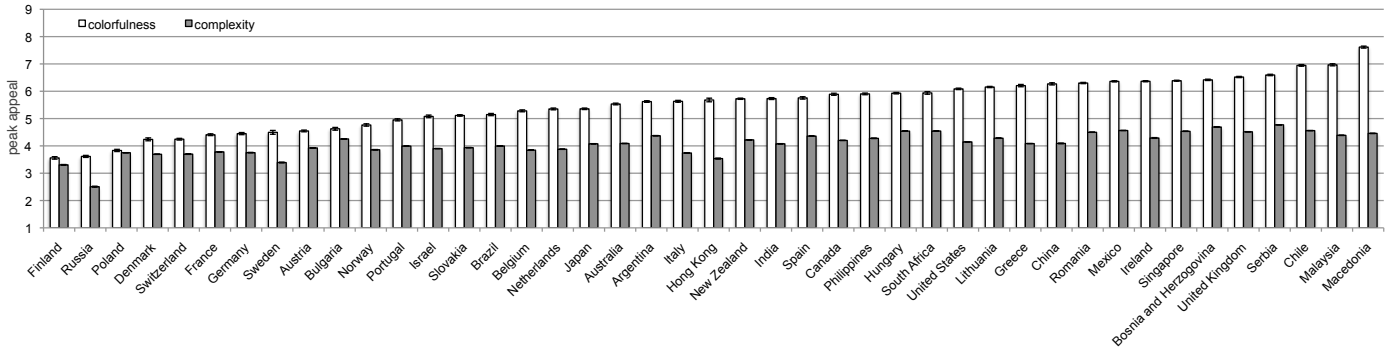


Figure 6. Colorfulness and complexity scores of peak appeal for different countries. Mean and standard errors for the scores of peak appeal are based on bootstrap resampling and curve-fitting of mean appeal ratings after applying locally weighted scatterplot smoothing (Lowess).

A significant interaction between colorfulness and country ($F_{(32)} = 75.89, p < .001$) further suggests that a preference for different levels of colorfulness is highly influenced by a person's country of residence. In most countries, ratings of appeal increase up to a moderate to high colorfulness before steeply declining (cf. Figure 5). However, for countries such as Finland, Russia, France, or Germany appeal peaks at a lower colorfulness than for most other countries, and steadily declines thereafter (see Figure 5). A comparison of the peaks of appeal shows that participants from these countries rated websites with a low colorfulness score (between 3.6 and 4.4, $SE = 0.04$) higher on appeal than participants from, for example, the United States (peak appeal=6.1, $SE = 0.03$). This is consistent with previous work suggesting that German websites use a smaller range of colors than the US [7]. Participants from Macedonia, Malaysia, and Chile have the highest preferences for colorful websites with their peak appeal ranging between 6.9 (for Chile and Malaysia) and 7.6 (for Macedonia, all $SE < 0.04$).

Education

People's preferences for colorfulness varies depending on education level ($F_{(7)} = 113.04, p < .001$). Independent of age, a lower education level indicates a higher preferences for colorful websites, and vice versa (as indicated by the beta statistics in the regression model output).

As Figure 7 illustrates, participants with a pre-high school education prefer websites with the highest colorfulness (peak appeal=6.9, $SE = 0.03$). Their ratings drop significantly for websites with a lower colorfulness (Cohen's d between low colorfulness vs. peak appeal = 1.7). Those who completed a doctoral education, in contrast, gave highest ratings for colorfulness levels of 5.0 ($SE = 0.04$).

We also observed a significant interaction between education level and visual complexity ($F_{(7)} = 28.80, p < .001$) with lower education levels preferring more complex websites. However, similar to our results for country and gender, complexity is less influential for appeal. Peaks range between 4.7 for people with pre-high school education ($SE = 0.01$) and 4.0 for people holding a PhD ($SE = 0.01$).

SUMMARY AND DISCUSSION

There are several important results. First, we found strong differences in first impressions between subgroups of varying age, gender, geography, and education. This was expected in view of the heterogeneous sample population, but it underlines that appeal is largely subjective even after a short exposure time of 500ms. The finding challenges previous assumptions that individual variability in aesthetic taste is small during the initial visceral reaction toward a design [22]. Second, our results showed that these differences in appeal can be partly explained by demographic background. This indicates similar preferences within specific subgroups of the same gender, country, education level, or similar age. In particular, we found that demographic background significantly affects preferences for colorfulness and complexity. Third, we identified how aesthetic preferences differ between these subgroups by providing the first ranking of complexity and colorfulness levels resulting in highest appeal by age group, gender, country, and education level, and pointing out several example websites that led to significant disagreements.

Our findings confirmed a number of previous results, such as that complexity is more important as a predictor of appeal than colorfulness [25], and that websites with a high visual complexity are generally disliked more than those with a low to medium complexity [8, 25, 17, 30]. We also substantiated the finding of Palmer and Schloss [23] that adults prefer more saturated (and thus, colorful) websites.

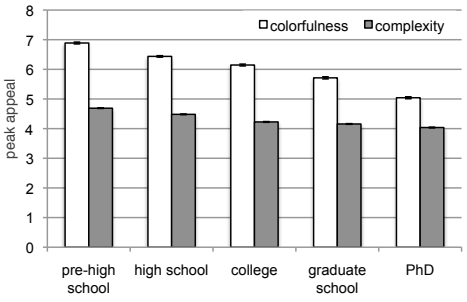


Figure 7. Colorfulness and complexity scores resulting in peak appeal for different education levels. Mean and standard errors of peal appeal based on bootstrap resampling and curve-fitting of mean appeal ratings after applying locally weighted scatterplot smoothing (Lowess).

We extended this previous knowledge with a more detailed account of how demographic background influences people's aesthetic preferences. For each subgroup, we presented estimates of the colorfulness and complexity levels resulting in

highest appeal. These peak estimates provide measurable evidence of differences in visual appeal. For example, while the average peak of appeal across all participants was found to be at a colorfulness score of 6.1 (on a scale from 1 to 9), we calculated that the peak appeal can range between 3.6 and 7.6 depending on a person's demographic background. The complexity levels of peak appeal differed less (peaks for different subgroups ranged between 2.5 and 4.8), but there are strong variations in people's tolerance for a low and a high visual complexity.

Analyzing the preferences of various subgroups, we found that females liked colorful websites more, and colorless websites less, than males. For complexity, both genders reached their peak appeal at a similar low to moderate complexity level, but females disliked simple websites more. Adults aged 41 years and above liked websites with a higher colorfulness and complexity than younger age groups. We also found a negative correlation between education level and colorfulness, as well as between education level and complexity. Independent of age, highly educated users prefer less complex and less colorful websites than others.

A user's geographical location is an additional factor influencing appeal. Many of our results on the varying preferences among people from different countries were unprecedented, but were nevertheless intuitive. In particular, we observed that countries in close proximity seem to share similar preferences. For example, the neighboring countries Finland and Russia preferred the lowest visual complexity and colorfulness of all countries. Participants from Macedonia, Serbia, and Bosnia and Herzegovina—all countries that were part of former Yugoslavia—had very similar preferences for highly colorful websites. In addition, the Northern European countries in our dataset (e.g., Denmark, Switzerland, France, Germany, Sweden, Austria) preferred a lower colorfulness than Southern European countries, such as Italy, Spain, Greece, or Romania. Northern European countries also preferred a lower colorfulness than Asian countries, such as China, Singapore, and Malaysia. Interestingly, Hong Kong and Japan preferred a lower colorfulness than other Asian countries (but higher than Northern European countries). All of the English-speaking countries Australia, New Zealand, Canada, United States, Ireland, and United Kingdom also preferred a higher colorfulness than Northern European countries. The results suggest that countries with a regular exchange of (cultural) values, e.g., due to migration, share similar website preferences.

With these results, we have demonstrated that people's aesthetic preferences for the design of websites can substantially differ. If the goal is to maximize website appeal, users should therefore receive designs personalized to their visual preferences. Our work takes a step towards this goal by contributing a list of objectively measured colorfulness and complexity levels of peak appeal for diverse demographic groups. In combination with the computational models introduced in [25], web designers can use these findings to estimate whether a website corresponds to what specific target groups will find most appealing.

LIMITATIONS AND FUTURE DIRECTIONS

Many of our results left us wondering about the underlying reasons for people's aesthetic preferences. For example, why did older participants prefer more complex websites? Here we can only speculate about possible causes, such as that the higher ratings for more complex sites are a generational effect; older users might simply be more used to text-heavy "web 1.0" designs. Determining causal relationships will be an important piece of future work.

A limitation of this paper is that it does not report on interactions between demographic variables. For example, our analysis showed that females liked colorful websites more than males, but that $>$40 year old females and those with a high education level (independent of age) prefer a similar colorfulness to the average male. While our model adequately incorporates these interactions, we leave a comprehensive report on these effects for future work.

In addition, any work on aesthetics risks loss of external validity due to the highly subjective nature of visual appeal. Although we analyzed a larger and more diverse sample than others, our work is no exception. To achieve our long-term goal of knowing what any user likes, we are therefore continuing to collect data online and are in the process of translating the study into a number of languages. We would be particularly enthusiastic to see our work replicated or extended with populations that are not online and have not yet been influenced by "global" websites.

Lastly, we are only beginning to understand how the exposure to, and identification with, certain demographic and cultural subgroups influences our visual preferences. Our hope is that our dataset will contribute to this knowledge by enabling a number of future analyses. For example, we collected additional demographic information based on plausible assumptions that migration, native language, the uptake of foreign languages, or a person's occupation influences appeal. The available data therefore opens opportunities for tackling questions, such as "How does migration impact visual appeal?"; "To what extent does a shared language result in more homogeneous preferences between countries?"; or "What is the influence of a mostly artistic occupation on design preferences?". We look forward to seeing these questions answered.

DATA SET

To enable replication and extension of our results, we make available the dataset and website stimuli: http://iis.seas.harvard.edu/resources/. We additionally provide an R-script detailing how variables were coded and entered into the regression, as well as the results of our final model and peak calculations.

ACKNOWLEDGMENTS

This work was funded in part by a Harvard Mind/Brain/Behavior Faculty Award and by a Sloan Research Fellowship. Katharina Reinecke was supported by the Swiss National Science Foundation under fellowship number PBZHP2-135922.

REFERENCES

1. Barber, W., and Badre, A. Culturability: The Merging of Culture and Usability. In *Conference on Human Factors & the Web* (1998).
2. Bates, D., Maechler, M., and Bolker, B. *lme4: Linear mixed-effects models using S4 classes*, 2013. R package version 0.999999-2.
3. Berlyne, D. *Studies in the New Experimental Aesthetics*. Washington, DC: Hemisphere Pub. Corp., 1974.
4. Bloch, P. Seeking the Ideal Form: Product Design and Consumer Response. *The Journal of Marketing 59*, 3 (1995), 16–29.
5. Chen, J. Y., Whitfield, T. W. A., Robertson, K., and Chen, Y. The Effect of Cultural and Educational Background in the Aesthetic Responses of Website Users. Tech. rep., National Institute for Design Research, Swinburne University of Technology, 2010.
6. Cyr, D., Head, M., and Larios, H. Colour Appeal in Website Design Within and Across Cultures: A Multi-method Evaluation. *Int. Journal of Human-Computer Studies 68*, 1-2 (2010), 1–21.
7. Cyr, D., and Trevor-Smith, H. Localization of Web Design: An Empirical Comparison of German, Japanese, and U.S. Website Characteristics. *Journal of the American Society for Information Science and Technology 55(13)* (2004), 1–10.
8. Geissler, G., Zinkhan, G., and Watson, R. The Influence of Home Page Complexity on Consumer Attention, Attitudes, and Purchase Intent. *Journal of Advertising 35*, 2 (2006), 69–80.
9. Gelman, A. Boot, 2013. http://andrewgelman.com/2013/06/03/boot/, last accessed September 16, 2013.
10. Hasler, D., and Suesstrunk, S. Measuring Colourfulness in Natural Images. In *Proc. SPIE/IS&T Human Vision and Electronic Imaging*, vol. 5007 (2003), 87–95.
11. Hoyer, W. D., and Stokburger-Sauer, N. E. The Role of Aesthetic Taste in Consumer Behavior. *Journal of the Academy of Marketing Science 40*, 1 (2011), 167–180.
12. Hsiu-Feng, W. Picture Perfect: Girls' and Boys' Preferences Towards Visual Complexity in Children's Websites. *Computers in Human Behavior* (2013).
13. Ivory, M., Sinha, R., and Hearst, M. Empirically Validated Web Page Design Metrics. In *Proc. CHI'01* (2001), 53–60.
14. Lindgaard, G., Dudek, C., Sen, D., Sumegi, L., and Noonan, P. An Exploration of Relations Between Visual Appeal, Trustworthiness and Perceived Usability of Homepages. *ACM ToCHI 18*, 1 (2011).
15. Lindgaard, G., Fernandes, G., Dudek, C., and Brown, J. Attention Web Designers: You Have 50 Milliseconds to Make a Good First Impression! *Behaviour & Information Technology 25*, 2 (2006), 115–126.
16. Martindale, C., Moore, K., and Borkum, J. Aesthetic Preference: Anomalous Findings for Berlyne's Psychobiological Theory. *The American Journal of Psychology* (1990), 53–80.
17. Michailidou, E., Harper, S., and Bechhofer, S. Visual Complexity and Aesthetic Perception of Web Pages. *Proc. Design of Communication* (2008), 215–224.
18. Moshagen, M., Musch, J., and Göritz, A. S. A Blessing, not a Curse: Experimental Evidence for Beneficial Effects of Visual Aesthetics on Performance. *Ergonomics 52*, 10 (2009), 1311–1320.
19. Moshagen, M., and Thielsch, M. T. Facets of visual aesthetics. *Int. Journal of Human-Computer Studies 68*, 10 (2010), 689–709.
20. Moss, G., and Gunn, R. Gender Differences in Website Production and Preference Aesthetics: Preliminary Implications for ICT Education and Beyond. *Behaviour & Information Technology 28*, 5 (2000), 447–460.
21. Nakagawa, S., and Schielzeth, H. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution 4* (2013), 133–142.
22. Ortony, A., Norman, D. A., and Revelle, W. *The role of affect and proto-affect in effective functioning*. Oxford University Press, 2005.
23. Palmer, S., and Schloss, K. Ecological Valence and Human Color Preferences. *Proc. National Academy of Science 107* (2010), 8877–8882.
24. Reinecke, K., and Bernstein, A. Improving Performance, Perceived Usability, and Aesthetics with Culturally Adaptive User Interfaces. *ACM ToCHI 18*, 2 (2011).
25. Reinecke, K., Yeh, T., Miratrix, L., Mardiko, R., Zhao, Y., Liu, J., and Gajos, K. Predicting Users' First Impressions of Website Aesthetics With a Quantification of Perceived Visual Complexity and Colorfulness. In *Proc. CHI'13* (2013).
26. Rosenholtz, R., Li, Y., and Nakano, L. Measuring Visual Clutter. *Journal of Vision 7*, 2 (2007).
27. Tractinsky, N. *Visual Aesthetics*. The Interaction Design Foundation, 2013.
28. Tractinsky, N., Cokhavi, A., Kirschenbaum, M., and Sharfi, T. Evaluating the Consistency of Immediate Aesthetic Perceptions of Web Pages. *Int. Journal of Human-Computer Studies 64* (2006).
29. Tuch, A. N., Bargas-Avila, J. A., and Opwis, K. Symmetry and Aesthetics in Website Design: It's a Man's Business. *Computers in Human Behavior 26*, 6 (2010), 1831–1837.
30. Tuch, A. N., Presslaber, E., Stoecklin, M., Opwis, K., and Bargas-Avila, J. The Role of Visual Complexity and Prototypicality Regarding First Impression of Websites: Working Towards Understanding Aesthetic Judgments. *Int. Journal of Human-Computer Studies 70(11)* (2012).
31. Zheng, X., Chakraborty, I., Lin, J., and Rauschenberger, R. Correlating Low-level Image Statistics with Users' Rapid Aesthetic and Affective Judgments of Web Pages. In *Proc. CHI'09* (2009).