International Journal of User-System Interaction 16 (2) 2023, 25-48

A Factorial Analysis of Visual Profiles of Websites on Mobile Devices

Nicolas Burny[1] and Luis A. Leiva[2]

[1]Louvain Research Institute in Management and Organizations (LouRIM), Université Catholique de Louvain
Place des Doyens, 1 – B-1348 Louvain-la-Neuve (Belgium)
E-mail: nicolas.burny@uclouvain.be

[2]Department of Computer Science, Faculty of Science, Technology and Medicine, University of Luxembourg
Avenue de la Fonte, 6 – L-4364 Esch-sur-Alzette (Luxembourg)

Abstract. Evaluating a user interface often means comparing it against a reference, whether theoretical or empirical. To ensure an appropriate comparison, this paper contributes to computational evaluation by creating a large dataset (top 53 websites ranked in 20 categories: N=1060), computing 53 metrics for them (56,180 entries) by applying a principal component analysis to identify standout metrics, and by performing a factorial analysis to derive a reference visual profile for websites within each category. After analysing the correlations between these 53 metrics and performing various rotations, we have been able to reduce the expressiveness of these metrics to seven latent factors: colorfulness, color complexity, analogous color scheme dissimilarity, primary hue, lightness deviation, complementary colors dissimilarity, and structural simplicity. Therefore, positioning a screenshot with respect to the corresponding visual profile can be simplified based on these latent factors. This approach aims to provide a more comprehensive and context-specific understanding of computational evaluation in the realm of graphical user interface design.

Keywords: Computational evaluation, Computational interaction, Factorial Analysis, Graphical User Interfaces, Principal component analysis, Visual design, Visual Profiles.

DOI: 10.37789/ijusi.2023.16.2.1

1. Introduction

Computational interaction (Oulasvirta et al., 2018) seeks to employ computational methods and techniques to reason about the structure of interactions, aiming to inform and improve the design of user interfaces.

Within this field, computational evaluation (Camargo et al., 2018) and modeling (Bauerly and Liu, 2006) employs metrics (Oulasvirta et al., 2018; Zen and Vanderdonckt, 2014), models (Leder et al., 2004; Ngo et al., 2003), principles (Lara-Alvarez and Reyes, 2017), techniques (Dondis, 1974; Vanderdonckt and Gillo, 1994) and laws to objectify user interface quality, ensuring and optimizing it (Camargo et al., 2018). A substantial amount of research has been dedicated to identifying, defining, calculating, and testing metrics (Burny and Vanderdonckt, 2022; Lara-Alvarez and Reyes, 2017; Moshagen et al., 2009; Reinecke et al., 2013; Skulmowski et al., 2016; Tuch et al., 2012) (also known as measures: Ngo et al., 2003) to characterize the overall quality of a user interface and its associated quality factors. Many of these metrics focus on assessing the visual quality of a user interface (Hartono and Holsapple, 2019; Lepreux et al., 2006), primarily its aesthetics (Bauerly and Liu, 2008; Leder et al., 2004; Wang et al., 2018; Zheng et al., 2009), but not exclusively, given its close correlation with quality factors, such as perceived usability in general (Tractinsky et al., 2000) and in model-based design (Dupuy-Chessa et al., 2016; Sousa et al., 2008), performance (Sonderegger and Sauer, 2009), credibility (Robins and Holmes, 2008), and trustwhortiness (Skulmowski et al., 2016). However, these metrics vary widely in their theoretical or empirical foundations, relying on different formulas, leading to erroneous and inconsistent interpretations.

Numerous software tools, such as WebTango (Ivory and Hearst, 2002), AIM (Oulasvirta et al., 2018), QUESTIM (Zen and Vanderdonckt, 2014), PLAIN (Soui et al., 2017), UI-CAT (Riegler and Holzmann, 2015), Web UI ANALYZER (Bakaev et al., 2019), have emerged to swiftly and efficiently calculate these metrics for graphical user interfaces like websites (which is convenient because the source code is in principle accessible) and mobile applications (which is convenient because they can be downloaded from an application store and there are datasets available), but less so those of professional applications. These tools, whether online (Zen and Vanderdonckt, 2014), offline (Soui et al., 2017), in the form of callable services (Riegler and Holzmann, 2015) or both (Oulasvirta et al., 2018), provide an estimated value for each metric, making the computational evaluation context-agnostic when reference values of these metrics are absent and context-sensitive when compared to known reference values. Some studies calculated these measures for a small number of user interfaces (Dupuy-Chessa et al., 2016), or a large number (Wang et al., 2018), but their

applicability remains questionable by the limited consideration given to the context of use and application domain: the metric values obtained for an online newspaper should not be compared to the same reference values as one for a gaming application. Comparing the calculated values with reference values for a large number of user interfaces is useful, but not context-specific. Other studies have calculated these measures instead for a significant number of interfaces within a given application domain (Camargo et al., 2018). However, applicability remains limited, as the problem of generalization arises: either the reference values are too specific to one domain and cannot be transposed to another, or the reference values are too generic and have little or no application to a given domain. For example, there are models based on machine learning or applying other methods to characterize visual design or aesthetics (Wang et al., 2018), but the problem of generalization/specialization persists. We do not know to what extent we can apply such a model and how to interpret it.

To address the major challenges mentioned above, this paper contributes to computational evaluation by initially creating a large dataset of mobile user interfaces (20 website categories with the top 53 sites ranked=1,060 websites). Subsequently, 53 metrics were systematically computed to augment the dataset to 1,060 × 53 values = 56,180 entries. Finally, principal component analysis is employed to identify standout metrics, and their distribution is analyzed through factor analysis to derive a reference visual profile for websites within each category. This approach aims to provide a more comprehensive and context-specific understanding of computational evaluation in the realm of graphical user interface design.

To this end, this paper is structured as follows: Section 2 conducts a literature review targeting recent or representative studies on the visual design of Graphical User Interfaces (GUIs) and computation of related metrics; Section 3 describes the experimental setup and provides a first analysis of the metrics on the dataset; Section 4 performs an exploratory data analysis of the metrics; Section 5 identifies the latent structure in the data and establishes the visual profile of website categories based on the extracted factors. Finally, Section 7 concludes this study by discussing its contributions, its limitations, and the potential improvements for the future.

2. Related Work

To assess GUI quality and associated factors (ISO/IEC 25010, 2011), extensive research has focused on defining and computing metrics (Moshagen et al., 2009; Ngo et al., 2003; Reinecke et al., 2013; Skulmowski et al., 2016; Touch et al., 2012), mostly emphasizing visual quality and aesthetics, closely linked to factors like usability or performance (Leder et al., 2004; Robins and Holmes, 2008; Tractinsky et al., 2000). However, these metrics vary widely in theoretical or empirical foundations, leading to potential misinterpretations or inconsistent results. Studies often calculate metrics for a given number of GUIs to discover any correlation between GUI features and quality dimensions or to develop models predicting the user score for a particular user experience facet (Ivory and Hearst, 2002; Reinecke et al., 2013; Zheng et al., 2009).

Ngo et al. (2003) presented a theoretical approach to quantifying GUI aesthetics through metrics computable by formulas to demonstrate a close relationship between them and perceived aesthetics. Based on Ngo's work, Zheng et al. (2009) studied low-level image features and measured their impact on user perceptions of visual design aesthetics. Ivory and Hearst (2002) came to the same conclusion after performing a quantitative analysis of web page layout and usability through 157 low-level measures computed on 5,300 web GUIs. The greatest interest of this study lies in its ability to evaluate any GUI by comparing the values calculated through the metrics with the reference values obtained for the corresponding website category, which is the hallmark of computational evaluation. When these reference values are recorded in a visual profile for each category, real benchmarking can be carried out. Evaluating typically refers to comparing a current GUI to something, like reference values of metrics, such as for comparing the consistency between two GUIs (Burny and Vanderdonckt, 2022). We will pursue the same goal in this paper with an up-to-date dataset, different metrics, and two statistical instruments. Others have attempted to capture such "visual profiles", but in rather different ways. Moshagen and Thielsch (2013) measured the perceived visual aesthetics of websites by creating the Visual Aesthetics of Websites inventory (VisAWI), of which a short version called VisAWI-S was developed and evaluated in three studies comprising 1673 participants. VisAWI-S is a reliable metric that captures a single dimension of perceived visual aesthetics and provides a good approximation

to the full-length version.

Bauerly and Liu (2006) investigated the effect of GUI elements on aesthetic evaluation and reported a strong correlation between visual elements and aesthetic ratings. Furthermore, the importance of symmetry and the number of compositional elements was demonstrated by measuring the effect of these two parameters on subjective appraisals of interface aesthetics (Bauerly and Liu, 2008). Robins and Holmes (2008) showed that the same GUI content with higher aesthetics is perceived as more credible. Tuch et al. (2012) also examined the impact of design factors on perceptions of visual appeal and found that low visual complexity and high prototypicality were perceived as highly appealing. This supports the information-processing stage model of aesthetics processing (Leder et al., 2004).

Based on low-level features of GUIs, Reinecke et al. (2013) predicted the initial impression of aesthetics based on perceptual models of color and visual complexity. Seckler et al. (2015) examined the relationship between objective aesthetic design factors and subjective aesthetic perceptions to conclude that the combination of high symmetry, low complexity, blue hue, medium brightness, and medium to high saturation resulted in the highest subjective scores. Other factors such as credibility or trustworthiness were also influenced. The results obtained by those studies are sometimes inconsistent (Seckler et al., 2015). In sum, a significant body of knowledge has investigated how GUI metrics are used to evaluate or predict quality factors, but how to evaluate another GUI that was not in each dataset remains open: results obtained for one study are not necessarily generalizable. Miniukovitch and de Angeli (2015) determined a computational model for measuring GUI aesthetics for desktop applications and introduced some metrics that can be reused for mobile GUIs.

3. Method

Our study pursues two goals: (1) to identify the latent factors underlying the low-level features of mobile website GUIs quantified through measures and (2), to establish the visual profile of top rank websites based on the identified factors.

3.1. Data Collection and Dataset Acquisition

Many studies focused on GUI for websites browsed in desktop configurations or mobile applications. This may not be representative of the whole ecosystem as most of the websites are browsed using mobile devices nowadays (Mittal and Mattela, 2019; Qazi et al., 2020). Some studies are also relying on datasets composed of arbitrarily selected GUIs. As a consequence, some categories of websites may not be sufficiently represented or even not present at all in the selected panel of websites. The understanding of low-level GUI features is more complex and their usage makes the derived insights less actionable.

In contrast, we collected data from the 53 most popular websites in each of the 22 domain categories provided by SimilarWeb via its SimilarWeb API. We employed UILAB (Burny and Vanderdonckt, 2021) to create one gallery to contain the websites of each category. The screenshots were automatically captured by a script with a resolution of 414 pixels (width) by 732 pixels (height), with a Device Pixel Ratio (DPR) of 3, a resolution of a high-end mobile device (Mittal and Mattela, 2019). Some screenshots had to be captured manually and some websites were removed due to request timeout or regional restrictions. After processing, removing outliers, and clustering, we included 53 websites × 20 categories = 1,060 websites in the dataset. We used UILAB to compute a subset of 53 AIM metrics (Oulasvirta et al., 2018) on each screenshot (see Table 1). These metrics cover color perception (Hasler and Süsstrunk, 2003; Miniukovich and de Angeli, 2015) and perceptual fluency (Balinsky, 2006; Wong, Carpendale, and Greenberg, 2003). In particular, each screenshot image is analysed with respect to all color harmonic templates (Cohen-Or et al., 2006) on the hue wheel (Fig. 1): when the colors of a screenhot fall into the corresponding gray area, it is considered to be harmonic in terms of color harmonization. The templates may be rotated by an arbitrary angle: for example, the L type can be inverted.

We did not compute other metrics, such as accessibility metrics as they were not relevant ofr our study. The AIM metrics returning compound results were decomposed, thus resulting in 53 individual metrics for a final dataset consisting of 1060 screenshots × 53 metrics = 56,180 entries.

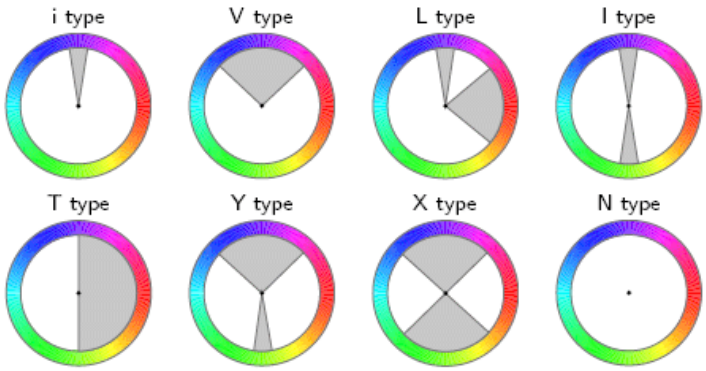

Figure 1. Color harmonic templates on the hue wheel (Hasler and Süsstrunk, 2003).

3.2. Exploratory Data Analysis

Fig. 2 shows that 53 metrics have markedly different univariate distributions: some measures have a clear bi-modal or multi-modal distribution, such as alpha-related variables associated with the angle at which the best fit occurs between the screenshot color scheme and the color scheme template represented by the variable. For example, Fig. 3 shows the results of color harmonization for the scrensnots of the whole distribution according to four templates: V, inverted L, L, and T. In particular, the rightmost graph shows that, with respect to the T template, the magenta spectrum falls in the grey area of the T type (Fig. 1) while the green spectrum, which is also dense, is outside the greay area, thus meaning that the distribution is somewhat harmonic with respect to this template.

Other metrics exhibit strong skewness due to statistical outliers, which might require removal for further analysis. Since all screenshots were manually verified, no other outlier was removed from the analysis.

Table 1. List of the 53 metrics computed for each web site on mobile device (Source: https://github.com/aalto-ui/aim/blob/aim2/metrics.json - Oulasvirta et al., 2018)

1	HSVAvg_average_hue	Hasler-Süsstrunk-Vetterli (HSV) color space, hue average
2	HSVAvg_average_sat	HSV color space, saturation average
3	HSVAvg_average_value	HSV color space, value average
4	HSVAvg_nb_unique_hsv	Number of distinct values of hue, saturation, and value in the HSV color space after color reduction; only values covering more than

		0.1% of image are counted
5	HSVAvg_nb_unique_hue	Number of unique hues in the HSV color space
6	HSVAvg_nb_unique_sat	Number of unique saturations in the HSV color space
7	HSVAvg_nb_unique_value	Number of unique values in the HSV color space
8	HSVAvg_sat_stdev	Number of saturations in HSV, standard deviation
9	HSVAvg_value_stdev	Number of values in HSV, value standard deviation
10	HasslerSusstrunk_Colorfulness	Hassler-Süsstrunk (HS) colorfulness in natural images
11	HasslerSusstrunk_RGYB_mean	Mean of red-green-yellow-blue values in HS
12	HasslerSusstrunk_RGYB_stdev	Standard deviation of red-green-yellow-blue values in HS
13	HasslerSusstrunk_RG_mean	Mean of red-green values in HS
14	HasslerSusstrunk_RG_stdev	Standard deviation of red-green values in HS
15	HasslerSusstrunk_YB_mean	Mean of yellow-blue values in HS
16	HasslerSusstrunk_YB_stdev	Standard deviation of yellow-blue values in HS
17	LABAvg_A_mean	LAB color space average of A metric
18	LABAvg_A_stdev	LAB color space standard deviation of A metric
19	LABAvg_B_mean	LAB color space average of B metric
20	LABAvg_B_stdev	LAB color space standard deviation of B metric
21	LABAvg_lightness_mean	LAB color space average of lightness
22	LABAvg_lightness_stdev	LAB color space standard deviation of lightness
23	Color_harmony_I_alpha	I template for the closest harmonic color scheme and distance to it: Harmonized image based on the closest harmonic template with minimal changes, significance level (alpha)
24	Color_harmony_I_distance	Distance to the I color harmonic template
25	Color_harmony_L_alpha	Significance level for L color harmonic template
26	Color_harmony_L_distance	Distance to the L color harmonic template
27	Color_harmony_Linverse_alpha	Significance level for inverted L color harmonic template
28	Color_harmony_Linverse_distance	Distance to the inverted L color harmonic template
29	Color_harmony_T_alpha	Significance level for T color harmonic template
30	Color_harmony_T_distance	Distance to the T color harmonic template
31	Color_harmony_V_alpha	Significance level for V color harmonic

		template
32	Color_harmony_V_distance	Distance to the V color harmonic template
33	Color_harmony_X_alpha	Significance level for X color harmonic template
34	Color_harmony_X_distance	Distance to the X color harmonic template
35	Color_harmony_Y_alpha	Significance level for Y color harmonic template
36	Color_harmony_Y_distance	Distance to the Y color harmonic template
37	Color_harmony_i_alpha	Significance level for i color harmonic template
38	Color_harmony_i_distance	Distance to the i color harmonic template
39	Edge_congestion	Mental effort needed to differentiate spatially proximal lines
40	Edge_density	Ratio of contour pixels to all pixels
41	FigureGround_contrast	Difference in color or luminance between two adjacent areas
42	Jpeg_file_size	File size in bytes of an screenshot image, saved in the JPEG format (image quality 70)
43	Luminance_stdev	Standard deviation of pixel luminance
44	Nb_alignment_lines	Number of alignment points
45	Nb_colors	Number of colors
46	Pixel_symmetry	Measure of symmetry in terms of pixels
47	Png_file_size	File size in bytes) of a screenshot image, saved in the PNG format (24-bit per pixel)
48	Quadtree_balance	Balance of the screenshot quadtree (tree that defines each node as having four children to subdivide a 2D space by splitting it recursively in four quadrants)
49	Quadtree_equilibrium	Equilibrium of the screenshot quadtree
50	Quadtree_nb_leafs	Number of leaf nodes of the screenshot quadtree
51	Quadtree_symmetry	Symmatry of the screenshot quadtree
52	Wave_score	Score of the Weighted Affective Valence Estimates (WAVE), defined as the mean of a mapping of pixel colors to the color preference values
53	White_space	Proportion of white space

The identification of latent factors is only reasonable if variables or groups of variables are correlated to some extent, indicating the presence of underlying concepts.

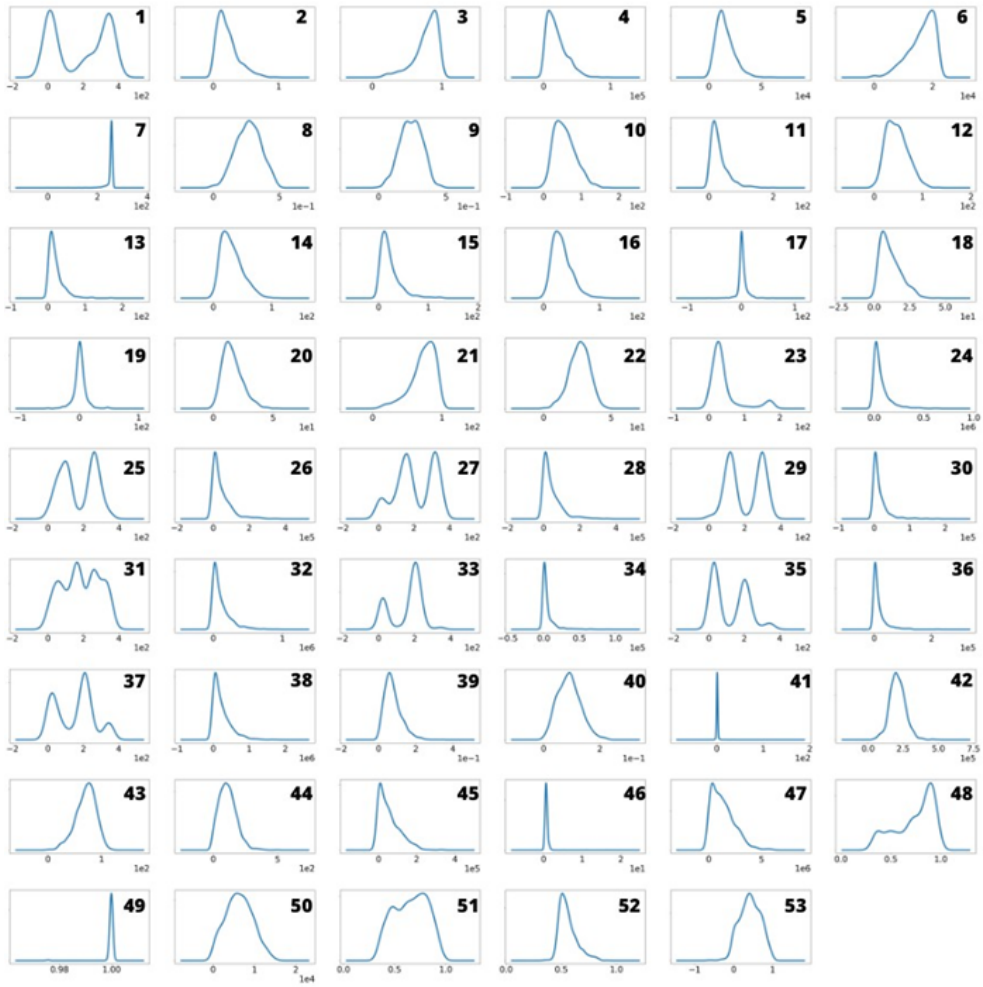

Figure 2. Statistical distribution of the 53 metrics computed on the whole dataset (see Table 1 for IDs).

Fig. 4 represents the correlation matrix between the 53 considered metrics, showing several clusters of highly correlated variables, which suggests that these data can be summarized as a smaller set of latent variables. To validate the data suitability for structure detection, we computed the Kaiser-Meyer-Olkin (KMO) (Kaiser, 1974) score for the whole dataset, which represents the proportion of variance in the variables that might be due to underlying factors: KMO=0.52, which is interpreted as “miserable”. After removing

variables with KMO scores lower than 0.6, the overall KMO score grew up to 0.8, which is interpreted as "meritorious", indicating the data are now suitable for factor analysis.

Moreover, we computed Bartlett's sphericity score for the dataset (Bartlett, 1951), a test that computes the probability that the correlation matrix has significant correlations among at least some of the variables in a dataset, a prerequisite for factor analysis. The test confirms the suitability of data for factorial analysis (B=54491.7, p <.001***). We ran a Principal Component Analysis (PCA) (Jolliffe and Cadima, 2016) to make a first selection of the number of factors to retain by applying the Kaiser criterion (Yeomans and Golder, 1982) to determine the number of components to keep for further analysis: the first 10 factors should be retained for a cumulated explained variance of 76.96%. Fig. 5 shows the Scree plot (Ledesma et al., 2015) representing the amount of variance explained by each extracted factor in decreasing order of eigenvalue.

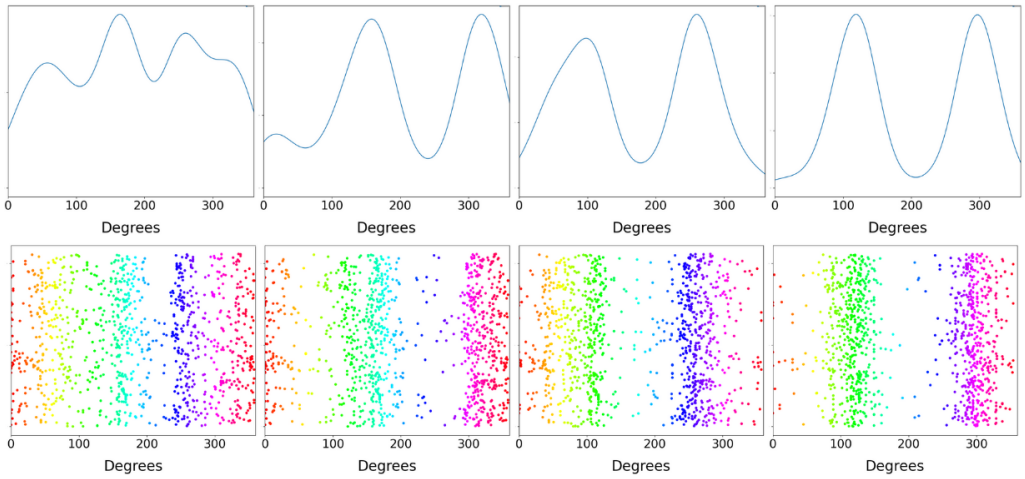

Figure 3. Screenshot metrics: densities and color distribution of variables according to four harmonic templates: color_harmony_V_alpha, color_harmony_Linverse_alpha, color_harmony_L_alpha, color_harmony_T_alpha.

4. Results and Discussion

4.1 Identification of Latent Factors

PCA is primarily used for feature summarization, data dimensionality reduction, and structure identification. However, interpreting principal components can be a challenging task, as they represent "generic" factors that load most of the variables and do not provide specific information about the data. Therefore, we utilized *Factor Analysis, specifically Exploratory Factor Analysis* (EFA) (Fabrigar et al., 1999), to discover underlying constructs and detect clusters of related variables, as it operates on the notion that observable variables can be reduced to a smaller set of hypothetical constructs. Initially, the appropriate factor space rotation method was determined to simplify the interpretation of the factor solution.

The ultimate aim of factor rotation is to achieve a simple structure (Tucker, 1955). To compare various factor extraction configurations, a plot of factor loadings (Fig. 6) was generated for 10 factors, including PCA and Factor analysis with PROMAX (an oblique rotation, which allows factors to be correlated), OBLIMIN (a method for oblique, nonorthogonal rotation), VARIMAX (an orthogonal rotation method that minimizes the number of variables that have high loadings on each factor), and QUARTIMAX (a rotation method that minimizes the number of factors needed to explain each variable) rotations, following recommendations of Brown (2009). We found that the oblique rotations PROMAX and OBLIMIN provide the simplest structure, while PCA without rotation suffered from cross-loadings and generic factors.

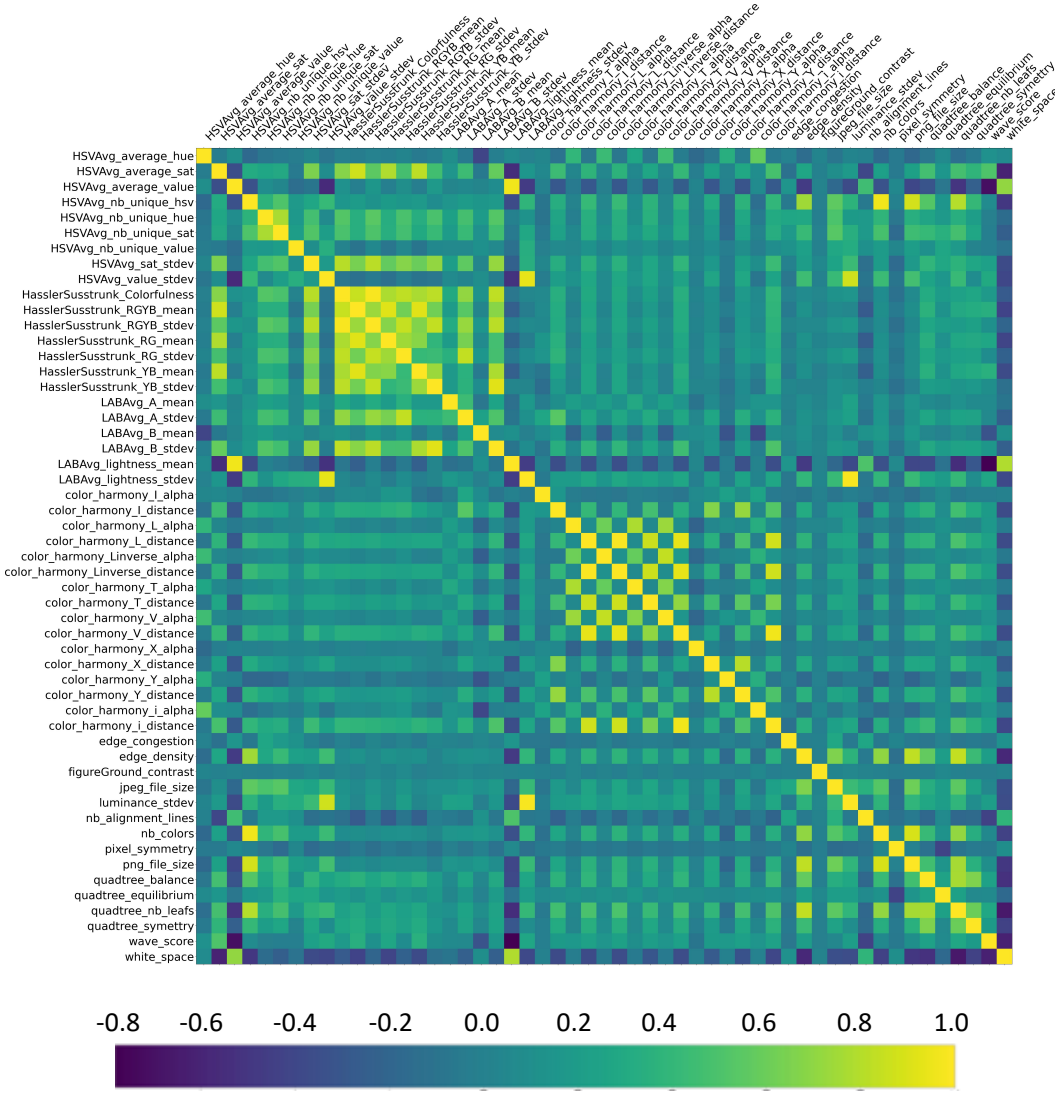

Figure 4. Correlation matrix between the 53 metrics computed in our dataset.

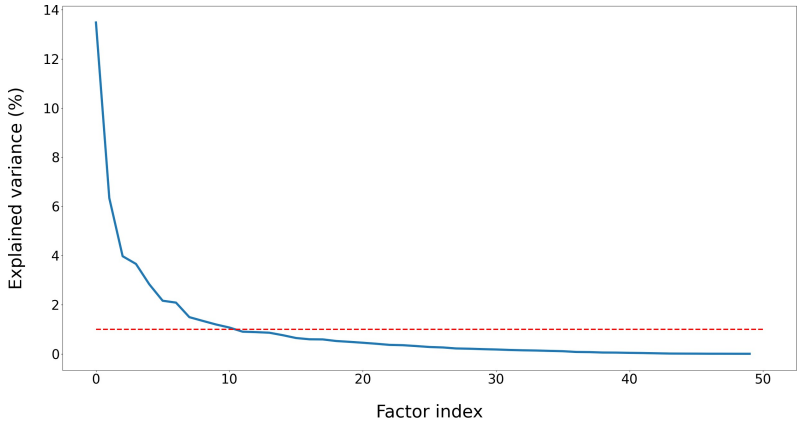

Fig. 5. Scree plot of Principal Component Analysis.

The decision to extract 10 factors was based on PCA, but this number may not be applicable to the oblimin rotation (Brown, 2009). Based on the Kaiser criterion and scree plot analysis, 8 factors with eigenvalues > 1 were retained, and parallel analysis confirmed that only the first 7 factors should be retained (Fig. 7). Due to the non-orthogonal nature of the destination space, oblique rotations allow extracted factors to be correlated with each other. Rotations shown in Fig. 7 are giving almost the same results as in Fig. 6, thereby suggesting that this ultimate reduction from 10 factors to 7 factors does not deteriorate the correlation between the variables.

Conversely, orthogonal rotations ensure that extracted factors are independent of one another. As part of the current scenario, oblique rotations were selected since the extracted factors represent distinct aspects of the same GUI, and a correlation between them is logical. Ultimately, the configuration with oblimin rotation was selected for interpretation purposes. The interpretation and labeling of extracted factors, the final seven latent factors of our study, are as follows:

1. **Colorfulness**: The most important loadings are all related to colorfulness and color variation, which corroborates previous results obtained by Hasler and Süsstrunk (2003) who developed this metric to assess colourfulness in natural images to perceptually qualify the effect that processing or coding has on colour. While this metric has been used to quantify the colourfulness of natural images, it is here used for the same

purpose on GUIs screenshots, which are assumed to have some natural dimension.

2. **Color complexity**: Several loadings are found to be associated with the number of colors present in the images. Furthermore, there is a positive correlation between the number of colors and the file size of both PNG and JPEG files, as the efficiency of image compression decreases with the number of colors. Additionally, the number of leaves in quadtree decomposition, as represented by the variable quadtree_nb_leafs, is also associated with the number of colors present in the images, since the decomposition process is based on color entropy, resulting in a larger number of leaves on average for images with a larger number of colors. Although the graph always breaks down into four nodes at each level, the number of levels is indicative of the visual complexity in terms of colors.
3. **Analogous color scheme dissimilarity**: This factor displays a positive correlation with the distance between the color scheme of an image and the analogous color templates (V, i) (Cohen-Or et al., 2006), as well as with the T and L Matsuda color schemes, which can also be categorized as analogous color schemes albeit with a higher degree of uncertainty. According to Lara-Alvarez and Reyes (2017), the Matsuda color schemes can be defined by three patterns, namely analog, complementary, and triad. Therefore, a higher value of this factor indicates a greater dissimilarity between the color scheme of the image and analogous color schemes, with a tendency towards a larger number of dissimilar colors, which may be categorized as complementary or triadic schemes.
4. **Primary hue**: This factor is associated with the hue angle at which the analogous color schemes (V , L, $L\_inverse$, T , i) are in closest proximity to the image color scheme. This factor can be interpreted as representing the principal color component or hue of the screenshot, which is further supported by the loading of HSVAvg_average_hue, because it is the most dominant and influential color, affecting both the visual impact (e.g., the

primary hue is often the most noticeable color in the screenshot if it covers the largest area of the screenshot or appears mostly in its elements, controls, etc.) and the psychological perception (e.g., such a hue sets the overall tone) of the screenshot.

5. **Lightness deviation**: The most significant factor loadings are associated with the deviation of lightness around the average value of lightness in the image. A higher value of this factor indicates a greater level of variation in luminosity observed in the screenshot. This factor contributes to creating a visual hierarchy of the elements in the screenshots by differentiating them by their lightness: for example, lighter or darker elementss can draw attention and guide the end user's eye to important elements of the screenshot, particularly the controls, such as pushbuttons or menu items.
6. **Complementary colors dissimilarity**: This factor is positively related to the distance to complementary color schemes (the greater the value of this factor, the greater the distance to complementary color schemes). It means that screenshots with large values for this factor are far from complementary schemes and tend to be closer to either analogous or triadic schemes. Complementary colors, being opposite each other on the color wheel, induce a better contrast among the screenshot elements, thus making it easier for the end user to identify and understand them and to recognize their type.
7. **Structural simplicity**: This factor is mostly related to the complexity of the image, but from a composition viewpoint (Lepreux et al., 2006) and a structural perspective (nb_alignment_lines representing grid quality of the image, white_space). The average lightness of the image LABAvg_lightness_mean is positively correlated with the proportion of white space and thus contributes to improving the structural simplicity. The higher the score, the simpler the GUI structure of the screenshot is because the white space is better distributed on the surface.

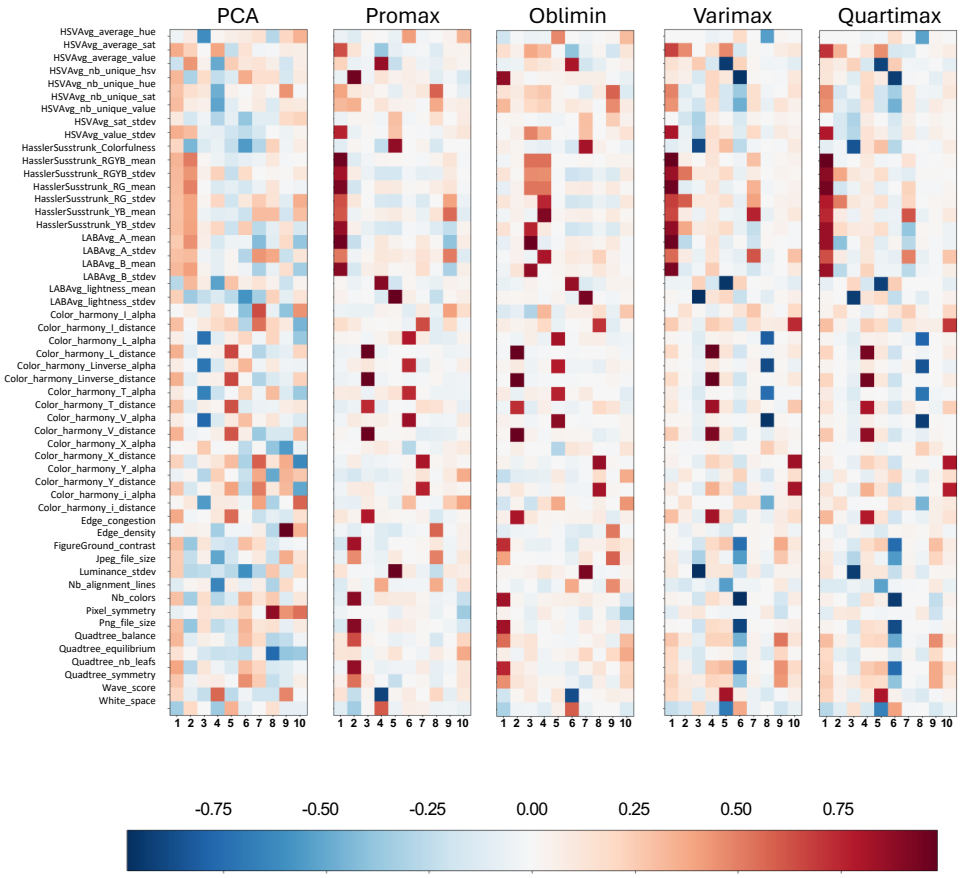

Fig. 6. Comparison of the different factorial analysis methods with 10 factors.

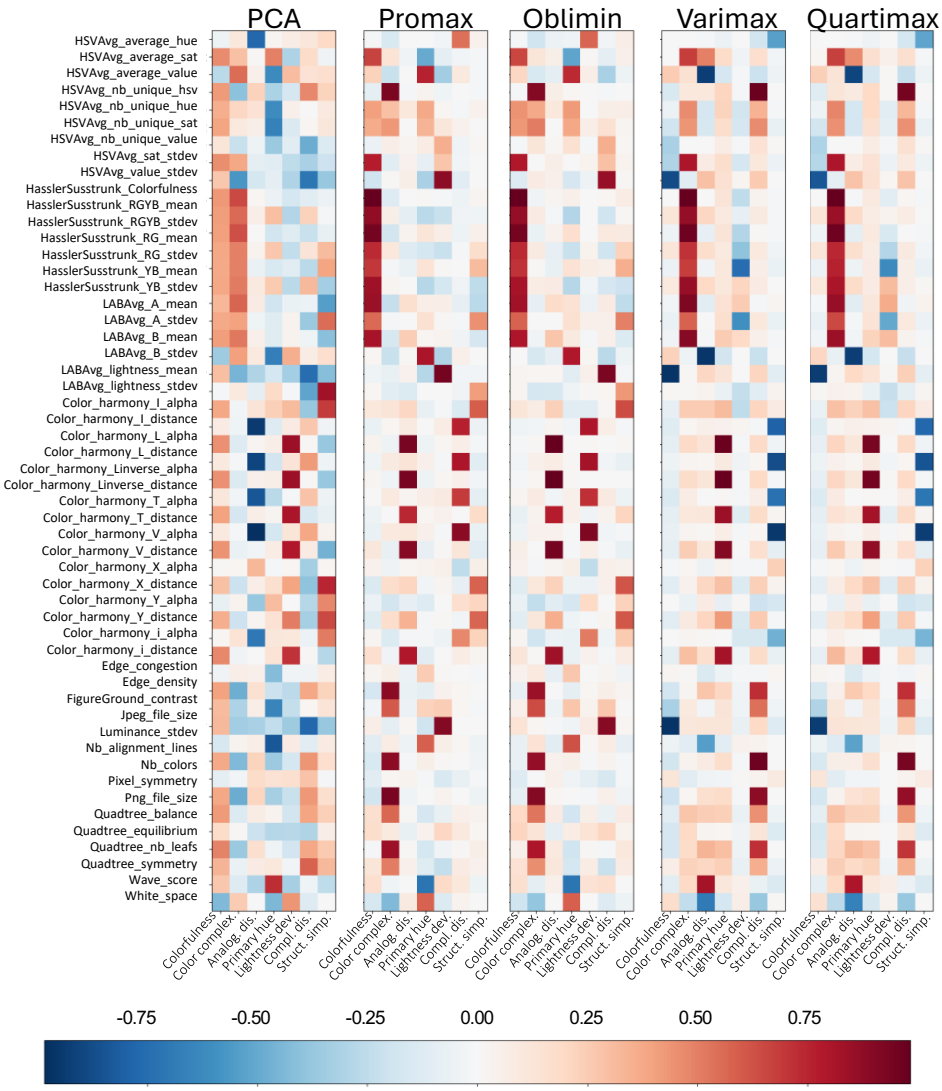

Fig. 7. Comparison of the different factorial analysis methods with 7 factors.

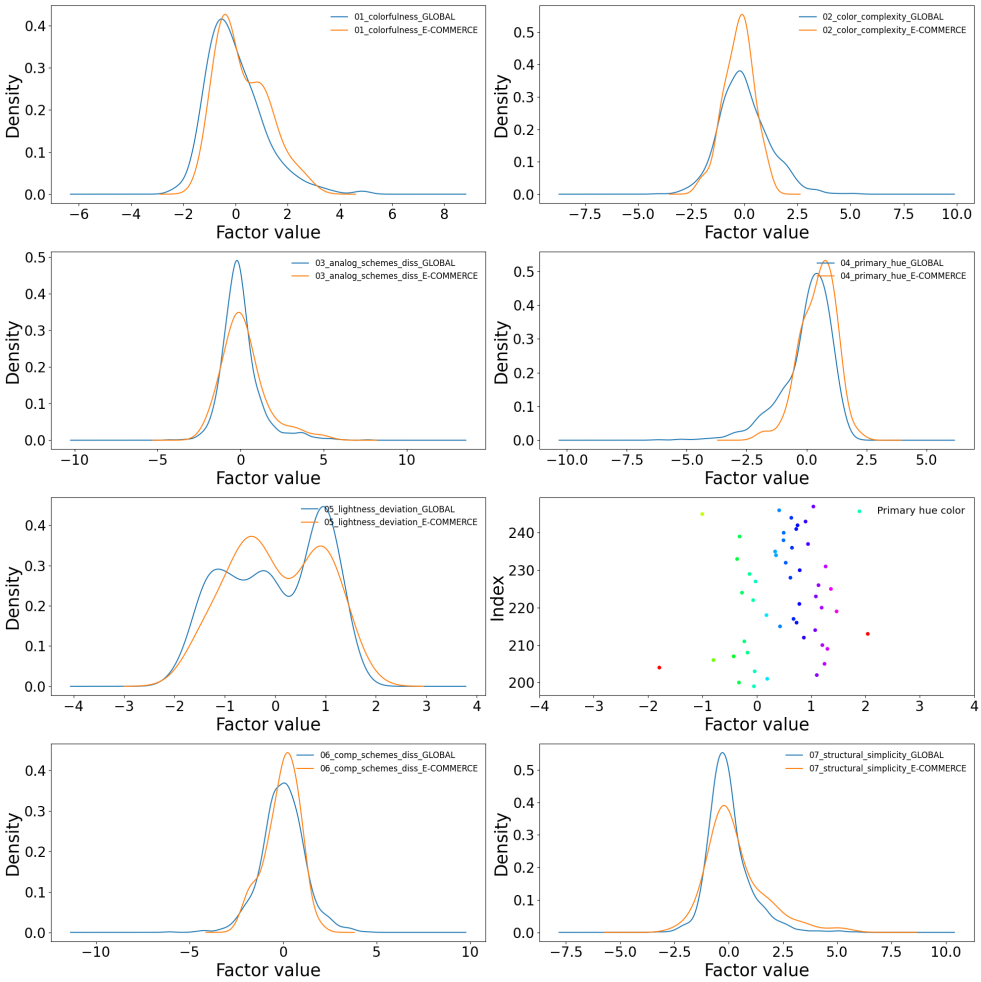

Fig. 8. Distribution of factors for the category E-commerce & Shopping.

4.2 Visual Profiles

To establish the visual profile of each website category, we first computed the average profile of websites by analyzing the distribution of factors across all categories. Subsequently, we compared the distribution of factors for each category with the global distribution of the same factors to identify significant deviations from the average profile.

To elaborate on the global distribution of factors, we observed that some factors were symmetrically distributed while others displayed a skewed distribution, either positively or negatively. Notably, the fifth factor, referred to as “primary hue”, mainly loads the alpha variables of several color schemes and displays a bimodal distribution, with peaks located around 120 and 300 degrees, corresponding to green and magenta colors, respectively. The minimum value of this distribution is located in the blue region of the hue wheel. Although the approximation of the two peaks of the distribution is not as precise as for single variables, owing to the superimposition of several alpha variables and other variables with a minor impact, it provides an insightful understanding of the fifth factor distribution.

As an example, we provide hereafter the analysis of a particular category of websites, namely “E-commerce and Shopping”: Fig. 8 shows that the color complexity of e-commerce websites is, on average, lower compared to other website categories. This can be determined from the greater concentration of the distribution of this factor around zero compared to the average profile distribution. Additionally, the distribution of the primary hue factor in this category is closer to a bi-modal distribution. The standard deviation of the distribution concerning the distances to color schemes is also larger, indicating a higher degree of uncertainty in associating a specific color scheme to websites of this category. Another notable difference between this category and the average profile is related to the structural simplicity of e-commerce websites, which appear to be much simpler in structure. This is likely attributed to the quality of the grid used in most of the websites in this category, as is evident from the shift in distribution towards positive values in the related chart. Other categories can be described similarly.

5. Limitations & Future Work

Despite the potential of using EFA for evaluating GUIs and extracting high-level constructs from low-level measures, there are several limitations and future work that need to be addressed. One of the main limitations of our study is the limited set of low-level metrics (Oulasvirta et al., 2018) that are considered for factor extraction. Although the metrics were carefully chosen based on their relevance to GUI evaluation, there may be other measures that could provide additional insights and improve the validity of the extracted

factors. Therefore, future studies could explore a wider range of low-level or high-level metrics to increase the comprehensiveness of the extracted factors. The novel method introduced in this paper suggests a procedure to be repeated on other metrics.

Secondly, the sample size used in this research could be enlarged to be more representative of the target population. While the current sample size was adequate for the analysis conducted, a larger sample size would allow for more robust statistical analyses and would increase the generalizability of the findings to other populations, especially when we know that different user groups perceive web aesthetics differently (Leiva et al., 2023). Additionally, future studies could consider stratified sampling methods to ensure that the sample is representative of the target population.

Thirdly, only one smartphone resolution was considered for every website. One could investigate the impact of screen resolution on the studied metrics. Some screen resolutions are also more frequently used depending on the platform, ranging from a small smartphone and tablets to high-resolution screens. Our study demonstrated that AIM metrics are particularly relevant to evaluating visual appeal and complexity of websites with a 414 x 732 screen resolution, but are perhaps less appropriate for larger resolutions.

Finally, there is a potential for subjective bias in the interpretation of factor loadings and naming. While efforts were made to ensure that the interpretation of the factors was based on objective evidence, there may be some degree of subjectivity in the interpretation. Future studies could consider using a panel of experts to validate the interpretation of the factors and ensure that they are representative of the construct being evaluated. Additionally, the use of a more objective naming convention for the factors could increase the clarity and reproducibility of the findings.

6. Conclusion

This paper presented a novel method for evaluating GUIs based on the extraction of high-level constructs from low-level metrics computed on GUIs. In this way, another GUI to be evaluated can be compared to the visual profile of top websites belonging to the same category. Assuming that the metric values contained in the visual profile become reference values, a true

benchmarking can be operated. There are limitations to the traditional approach of evaluating GUIs solely based on low-level measures, as it does not provide designers and other stakeholders with actionable insights. This study demonstrates that exploratory factorial analysis can be used to extract high-level constructs from low-level metrics to produce more interpretable insights. Through this analysis, we were able to extract high-level constructs with greater interpretability than low-level variables.

There are, however, some limitations that need to be addressed in future research. One limitation is the limited set of low-level measures considered for factor extraction, and future studies could explore a wider range of low-level measures to increase the comprehensiveness of the extracted factors. Another limitation is the sample size used in the study, which could be larger to increase the generalizability of the findings. Finally, there is a potential for subjective bias in the interpretation of factor loadings and naming, and future studies could consider using a panel of experts to validate the interpretation of the factors and ensure that they are representative of the construct being evaluated. Overall, the proposed methodology shows promise for improving GUI evaluation, and future work could build on this research to further develop and refine the method.

Open science. We release a spreadsheet containing the 53 metrics computed for the 1060 web sites on mobile devices as an accompanying resource to this paper.

Acknowledgments

The authors are all supported by the EU EIC Pathfinder-Awareness Inside challenge "Symbiotik" project (1 Oct. 2022-31 Dec. 2026) under Grant no. 101071147.

References

Bakaev, M., Heil, S., Khvorostov, V., Gaedke, M. (2019). Auto-Extraction and Integration of Metrics for Web User Interfaces. *Journal of Web Engineering 17(3)*, 561–590. https://doi.org/10.13052/jwe1540-9589.17676

Balinsky, H. (2006). Evaluating Interface Aesthetics: Measure of Symmetry. *Digital publishing*, vol. 6076. SPIE, 52-63. https://doi.org/10.1117/12.642120

Bartlett, M.S. (1951). The Effect of Standardization on a 2 Approximation in Factor Analysis. *Biometrika 38(3-4)*, 337–344. https://doi.org/10.1093/biomet/38.3-4.337

Bauerly, M., Liu, Y. (2006). Computational modeling and experimental investigation of effects of compositional elements on interface and design aesthetics. *International Journal of Human-Computer Studies 64(8)*, 670–682.

https://doi.org/10.1016/j.ijhcs.2006.01.002

Bauerly, M.S., Liu, Y. (2008). Effects of Symmetry and Number of Compositional Elements on Interface and Design Aesthetics. *International Journal of Human-Computer Interaction 24(8)*, 275–287. https://doi.org/10.1080/10447310801920508

Burny, N., Vanderdonckt, J. (2021). UILAB, a Workbench for Conducting and Reproducing Experiments in GUI Visual Design. *Proc. ACM Hum.-Comput. Interact. 5*, Article 196. https://doi.org/10.1145/3457143

Burny, N. and Vanderdonckt, J. (Semi-)Automatic Computation of User Interface Consistency. *Proc. of ACM Symposium on Engineering Interactive Computing Systems (EICS '22)*. Association for Computing Machinery, New York, 2022, 5–13. https://doi.org/10.1145/3531706.3536448

Brown, J.D. (May 2009). *Choosing the Right Number of Components or Factors in PCA and EFA*. Shiken: JALT Testing & Evaluation SIG Newsletter, Vol. 13, No. 2, 19-23. https://teval.jalt.org/test/bro_30.htm

Camargo, M.C., Barros, R.M., Barros, V.T.O. Visual Design Checklist for Graphical User Interface Evaluation. *Proc. of the 33rd Annual ACM Symposium on Applied Computing (SAC '18)*. Association for Computing Machinery, New York, 2018, 670–672. https://doi.org/10.1145/3167132.3167391

Cohen-Or, D., Sorkine, O., Gal, R., Leyvand, T., Xu, Y.-Q. (2006). Color Harmonization. *ACM Transactions on Graphics 25(3)*, 624–630.

https://doi.org/10.1145/1141911.1141933

Dondis, D.A. A Primer of Visual Literacy. The MIT Press, MA, USA, 1974. https://mitpress.mit.edu/9780262540292/a-primer-of-visual-literacy/

Dupuy-Chessa, S., Laurillau, Y., Céret, E. Considering Aesthetics and Usability Temporalities in a Model Based Development Process. Proc. of 28th Conference Francophone sur l'Interaction Homme-Machine (IHM '16). Association for Computing Machinery, New York, 2016, 25–35. https://doi.org/10.1145/3004107.3004122

Fabrigar, L.R., Wegener, D.T., MacCallum, R.C., Strahan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods 4(3)*, 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Hartono, E., Holsapple, C.W. (2019). Website Visual Design Qualities: A Threefold Framework. *ACM Transactions on Management Information Systems 10(1)*, Article 1. https://doi.org/10.1145/3309708

Hasler, D., Süsstrunk, S.E. (2003). Measuring Colorfulness in Natural Images. *Human Vision*

and Electronic Imaging VIII, Vol. 5007. SPIE, 87-95. https://doi.org/10.1117/12.477378

Ivory, M.Y., Hearst, M.A. Statistical Profiles of Highly-Rated Web Sites. Proc. of the ACM Conference on Human Factors in Computing Systems (CHI '02). Association for Computing Machinery, New York, 2002, 367–374.

https://doi.org/10.1145/503376.503442

Jolliffe, I.T., Cadima, J. 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A 374(2065)*, 20150202–20150202. https://doi.org/10.1098/rsta.2015.0202

Kaiser, H. (1974). An index of factorial simplicity. *Psychometrika 39(1)*, 31–36. https://doi.org/10.1007/BF02291575

Lara-Alvarez, C., Reyes, T. (2017). A Geometric Approach to Harmonic Color Palette Design. *Color Research & Application 44(1)*, 106–114. https://doi.org/10.1002/col.22292

Leder, H., Belke, B., Oeberst, A., Augustin, D. (2004). A Model of Aesthetic Appreciation and Aesthetic Judgments. *British Journal of Psychology 95(12)*, 489–508. https://doi.org/10.1348/0007126042369811

Ledesma, R., Valero-Mora, P., Macbeth, G. (2015). The Scree Test and the Number of Factors: a Dynamic Graphics Approach. *The Spanish Journal of Psychology 18(11)*. https://doi.org/10.1017/sjp.2015.13

Leiva, L.A., Hota, A., Oulasvirta, A. Enrico: A Dataset for Topic Modeling of Mobile UI Designs. *Proc. of 22nd International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '20)*. Association for Computing Machinery, New York, 9:1–9:4. https://doi.org/10.1145/3406324.3410710

Leiva, L.A., Shiripour, M., Oulasvirta, A. (2023). Modeling how different user groups perceive webpage aesthetics. *Universal Access to Information Society 22(4)*, 1417–1424. https://doi.org/10.1007/S10209-022-00910-X

Lepreux, S., Vanderdonckt, J., Michotte, B. Visual Design of User Interfaces by (De)composition. *Proc. of International Workshop on Interactive Systems. Design, Specification, and Verification (DSVIS '06)*. Lecture Notes in Computer Science, Vol. 4323. Springer, Berlin, 2006, 157–170. https://doi.org/10.1007/978-3-540-69554-7_13

Miniukovich, A., de Angeli, A. (2015). Computation of Interface Aesthetics. *Proc. of the 33rd International ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, 1163-1172.

https://doi.org/10.1145/2702123.2702575

Mittal, S., Mattela, V. (2019). A survey of techniques for improving efficiency of mobile web browsing. *Concurrency and Computation: Practice and Experience 31(15)*, e5126. https://doi.org/10.1002/cpe.5126