



Development of measurement instrument for visual qualities of graphical user interface elements (VISQUAL): a test in the context of mobile game icons

Henrietta Jylhä¹ · Juho Hamari¹

Received: 20 February 2019 / Accepted in revised form: 28 March 2020
© The Author(s) 2020

Abstract

Graphical user interfaces are widely common and present in everyday human–computer interaction, dominantly in computers and smartphones. Today, various actions are performed via graphical user interface elements, e.g., windows, menus and icons. An attractive user interface that adapts to user needs and preferences is progressively important as it often allows personalized information processing that facilitates interaction. However, practitioners and scholars have lacked an instrument for measuring user perception of aesthetics within graphical user interface elements to aid in creating successful graphical assets. Therefore, we studied dimensionality of ratings of different perceived aesthetic qualities in GUI elements as the foundation for the measurement instrument. First, we devised a semantic differential scale of 22 adjective pairs by combining prior scattered measures. We then conducted a vignette experiment with random participant ($n=569$) assignment to evaluate 4 icons from a total of pre-selected 68 game app icons across 4 categories (concrete, abstract, character and text) using the semantic scales. This resulted in a total of 2276 individual icon evaluations. Through exploratory factor analyses, the observations converged into 5 dimensions of perceived visual quality: Excellence/Inferiority, Graciousness/Harshness, Idleness/Liveliness, Normalness/Bizarreness and Complexity/Simplicity. We then proceeded to conduct confirmatory factor analyses to test the model fit of the 5-factor model with all 22 adjective pairs as well as with an adjusted version of 15 adjective pairs. Overall, this study developed, validated, and consequently presents a measurement instrument for perceptions of visual qualities of graphical user interfaces and/or singular interface elements (VISQUAL) that can be used in multiple ways in several contexts related to visual human–computer interaction, interfaces and their adaption.

Keywords Measurement instrument · Questionnaire · Aesthetics · Design guidelines · Graphical user interfaces · Adaptive user interfaces

✉ Henrietta Jylhä
henrietta.jylha@tuni.fi

Extended author information available on the last page of the article

1 Introduction

Aesthetics considerations in computers and other devices have quickly started to garner attention as the means to positively affect usability and satisfaction (Ahmed et al. 2009; Maity et al. 2015, 2016; Norman 2004; Tractinsky et al. 2000). Studies have shown that a user interface with balanced elements promotes user engagement, while a cluttered interface may result in frustration (Jankowski et al. 2016, 2019; Lee and Boling 1999; Ngo et al. 2000; Salimun et al. 2010). Moreover, adaptation within user interfaces has been shown to lead into higher ratings in look and feel as well as long-term usage of platforms (Debevc et al. 1996; Hartmann et al. 2007; Sarsam and Al-Samarraie 2018). This reflects the well-established knowledge in product design and marketing: aesthetics matter (e.g., Hartmann et al. 2007; Tractinsky et al. 2000), and collaboration between artists and technologists is essential in this regard (Ahmed et al. 2009). Increasing demands for customization within human–computer interaction introduce new possibilities and challenges to designers, which justifies further research on the topic.

Graphical user interface (GUI) is a way for humans to interact with devices through windows, menus and icons.¹ User interaction is enabled through direct manipulation of various graphical elements and visual indicators (e.g., icons) that are designed to provide an intuitive representation of an action, a status or an app.² Graphical user interfaces are widely used due to their intuitiveness and immediate visual feedback. Several factors have influenced the tremendous progress that GUI design has seen, such as advances in computer hardware and software as well as industry and consumer demands. Moreover, user interfaces adapt to individual user preferences by changing layouts and elements to different needs and contexts. Hence, a user interface attractive to individual users is increasingly important for companies aiming to positively contribute to their commercial performance (Gait 1985; Lin and Yeh 2010).

Aesthetics in GUI design refers to the study of natural and pleasing computer-based environments (Jennings 2000). It extends across the definition of fonts to pictorial illustrations, transforming information into visual communication through balance, symmetry and appeal.

Attention to pure aesthetics in GUI design is important in sustaining user interest and effectiveness in a service (Gait 1985). However, it has been noted that prior research has mainly focused in usability, perhaps at the expense of visual aesthetics, although aesthetic design is an integral part of a positive user experience as well as user engagement (Ahmed et al. 2009; Kurosu and Kashimura 1995; Maity et al. 2015; Ngo et al. 2000; Overby and Sabyasachi 2014; Salimun et al. 2010; Tractinsky et al. 2000). Within the field of graphical user interfaces, appealing designs have proven to enhance usability (Kurosu and Kashimura 1995; Ngo et al. 2000;

¹ Linux Information Project, “GUI Definition,” <http://www.linfo.org/gui.html> (accessed October 23, 2018).

² Android Developers, “Iconography,” <http://www.androiddocs.com/design/style/iconography.html> (accessed October 15, 2018).

Salimun et al. 2010; Sarsam and Al-Samarraie 2018; Tractinsky 1997; Tractinsky et al. 2000) as well as sense of pleasure and trust (Cyr et al. 2006; Jordan 1998; Zen and Vanderdonck 2016). A positive user experience is essential for successful human–computer interaction, as a user quickly abandons an interface that is connected with negative experiences. As the user experience is increasingly tied to adaptive visual aesthetics, it motivates the need for further research on graphical user interface elements. Perceptions of successful (i.e., appealing) visual aesthetics are subjective (Zen and Vanderdonck 2016), which complicates creating engaging user experiences for critical masses. Theories and tools have been proposed to assess and design appropriate graphical user interfaces (e.g., Choi and Lee 2012; Hassenzahl et al. 2003; Ngo et al. 2000; Ngo 2001; Ngo et al. 2003; Zen and Vanderdonck 2016), yet no consensus exists on a consistent method to guide producing successful user interface elements considering the subjective experience. In the pursuit of investigating what aesthetic features appear together in graphical icons, we attempt to address this gap by developing an instrument that measures graphical user interface elements via individual user perceptions.

First, we devised a semantic differential scale of 22 adjective pairs. We then conducted a survey-based vignette study with random participant ($n = 569$) assignment to evaluate 4 icons from a total of pre-selected 68 game app icons across 4 categories (concrete, abstract, character and text) using the semantic scales. Game app icons were used for validity and comparability in the results. This resulted in a total of 2276 individual icon evaluations. The large-scale quantitative data were analyzed in several ways. Firstly, we examined factor loadings of the perceived visual qualities with exploratory factor analysis (EFA). Secondly, we performed confirmatory factor analyses (CFA) to test whether the proposed theory could be applied to similar latent constructs. Although further validation is required, the results show promise. Based on these studies, we compose VISQUAL, an instrument for measuring individual user perceptions of visual qualities of graphical user interface elements, which can be used for research into adaptive user interfaces. Therefore, this study allows for theoretical and practical guidelines in the designing process of personalized graphical user interface elements, analyzed via 5 dimensions: Excellence/Inferiority, Graciousness/Harshness, Idleness/Liveliness, Normalness/Bizarreness and Complexity/Simplicity.

2 Visual qualities of graphical user interfaces

2.1 Variations of user-adaptive graphical user interfaces

Graphical user interface design has experienced tremendous change during the past decades due to technological evolution. An increasing diversity of devices have adopted interfaces that adapt according to device characteristics and user preferences. An adaptive user interface (AUI) is defined as a system that changes its structure and elements depending on the context of the user (Schneider-Hufschmidt et al. 1993), hence the UI has to be flexible to satisfy various needs. User interface adaptation consists of modifying parts or a whole UI. User modeling algorithms in

the software level provide the personalization concept, while GUIs display the content, expressing personalization from the user's perspective (Alvarez-Cortes et al. 2009). For example, UI elements are expected to scale automatically with screen size and hide unwanted menu elements. Adaptation can be divided into two categories depending on the end user: adaptability and adaptivity. Adaptability means the user's ability to adapt the UI, and adaptivity means the system's ability to adapt the UI. When users communicate with interfaces, both the human and the machine collaborate toward adaptation, i.e., mixed initiative adaptation (Bouzit et al. 2017). Adaptiveness in interfaces has been widely studied in terms of user performance (Gajos et al. 2006), preference (Cockburn et al. 2007) and satisfaction (Gajos et al. 2006), as well as improving task efficiency and learning curve (Lavie and Meyer 2010).

The most important advantage of AUIs is argued to be the total control of UI appearance that the user has, although it is at the same time considered a shortcoming for users with lower level of technology experience and skill (Gullà et al. 2015). Adaptive user interfaces may in many cases result in undesired or unpredictable interface behavior because of the challenges in specifying the design for the wide variety of users which in some cases lead to users not accepting the UI (Alvarez-Cortes et al. 2009; Bouzit et al. 2017; Gajos et al. 2006). Moreover, prior research (Gajos et al. 2006) has shown that purely mechanical properties of an adaptive interface lead to poor user performance and satisfaction. Therefore, understanding user preferences and perceptions is essential in creating interfaces, and it is necessary to assess these in early stages of the design process to effectively identify different user profiles (Gullà et al. 2015). Due to the rapid changes to UI design, new adaptation techniques and systematic methods are needed in which design decisions are led by appropriate parameters concerning users and contexts.

2.2 Measuring visual qualities of graphical user interfaces

A distinction has been made between two types of aesthetics within human–computer interaction, namely classical and expressive aesthetics (Hartmann et al. 2008). Classical aesthetics refers to orderly and clear designs, whereas expressive aesthetics refer to creative and original designs. Classical aesthetics seem to be perceived more evenly by users, while expressive aesthetics are denounced by more dispersion depending on contextual stimuli (Mahlke and Thüring 2007). Aesthetic value of graphical user interfaces has been attempted to measure objectively by several geometry-related and image-related metrics, e.g., balance, equilibrium, symmetry and sequences well as color contrast and saturation to avoid human involvement in the process (Maity et al. 2015, 2016; Ngo et al. 2000, 2001, 2003; Vanderdonckt and Gillo 1994; Zen and Vanderdonckt 2014, 2016). These visual techniques in the arrangement of layout components can be divided into physical techniques, composition techniques, association and disassociation techniques, ordering techniques, as well as photographic techniques (Vanderdonckt and Gillo 1994). Furthermore, balance is defined as a centered layout where components are equally weighed. Equilibrium is defined as equal balance between opposing forces. Symmetry is defined

as the equal distribution of elements. Sequence is defined as the arrangement of elements in such a way that facilitates eye movement (Ngo et al. 2003). Color contrast is the difference in visual properties that distinguishes objects from each other and the background, while saturation indicates chromatic purity (Maity et al. 2015).

A user interface is said to be in a state of repose when all of these metrics are configured accordingly. Correspondingly, if these metrics are not perfected, it will result in a state of chaos (Ngo et al. 2000). Prior research has aligned these metrics with user perceptions (Maity et al. 2015; Ngo et al. 2000; Salimun et al. 2010; Zen and Vanderdonck 2016) and task performance (Salimun et al. 2010), which has led to inconsistent results. Initial findings (Maity et al. 2015; Ngo et al. 2000) report high correlations between computed aesthetic value and the aesthetics ratings of design experts, artists and users. These results were replicated only to an extent by a study (Zen and Vanderdonck 2016) that reported medium degree of inter-judge agreement and low reliability for calculating symmetry and balance, after which a new formula for balance is introduced. Another study (Salimun et al. 2010) computed several metrics based on the prior literature (Ngo 2001; Ngo et al. 2003) to conclude that some metrics, such as symmetry and cohesion, influence results more than others. A study (Möttus et al. 2013) that tested objective and subjective evaluation methods according to the prior literature (Ngo et al. 2000, 2003) displayed a weak correlation between the ratings.

In addition to metric-based instruments, aesthetic value of graphical user interfaces has been measured by empirical approaches (Choi and Lee 2012; Hassenzahl et al. 2003; Hassenzahl 2004). Focusing on facets of simplicity for smartphone user interfaces, Choi and Lee (2012) developed a survey-based method incorporating the following six components: reduction, organization, component complexity, coordinative complexity, dynamic complexity, and visual aesthetics. Results showed that the instrument was successful in predicting user satisfaction by simplicity perception (Choi and Lee 2012). A seven-point semantic differential scale was introduced by Hassenzahl et al. (2003) with 21 items measuring hedonic quality–identification, hedonic quality–stimulation, and pragmatic quality. The instrument was further tested by Hassenzahl (2004) with a version that included two evaluational constructs (ugly–beautiful and bad–good), resulting in 23 semantic differential items. Prior research investigated graphical user interfaces of MP3 software and found that beauty is related to hedonic qualities rather than pragmatic qualities (Hassenzahl 2004).

Prior literature (Maity et al. 2015, 2016; Zen and Vanderdonck 2016) suggests that contradictory results in metric-based evaluation theories and tools of aesthetics in GUI research are perhaps caused by analyzing user interfaces as entities without considering the content. This gap in calculating aesthetics with metric-based evaluations means that many metric evaluations consider a graphical user interface as a single piece although it essentially consists of different elements with specific purposes and designs (Maity et al. 2015). For instance, designing an interactive button is very different from defining type faces in that these elements serve different purposes in user interfaces (Maity et al. 2016). Moreover, empirical studies on GUI aesthetics have often relied on website layouts as study objects (Hassenzahl 2004). This can be problematic, as measuring perceived attractiveness of website layouts does

not necessarily reveal which elements in the user interface are successful. Layout designs vary, which may cause difficulties in generalization. This can be regarded as a shortcoming of the empirical measurements as inclusivity may prevent calculating genuine values of user interfaces. Prior study (Vanderdonckt and Gillo 1994) attempting to automate calculation of visual techniques with single interface components found that some techniques could be measured, such as physical techniques, while some others appeared more challenging to measure, such as photographic techniques. We note that contextual factors surrounding single GUI components are important in affecting user perceptions, thus evaluating GUI elements separately may in some cases prove challenging. Moreover, the application of principles heavily depends on visual aims, and hence, further comparison between measurement instruments is needed in order to explore the relationship between single components and their context.

In order to address these gaps, and rather than experimenting with a graphical user interface as a single piece, we scaled the validation of VISQUAL into single interface components, i.e., icons. Icons are pictographic symbols within a computer system, applied principally to graphical user interfaces (Gittins 1986) that have replaced text-based commands as the means to communicate with users (García et al. 1994; Gittins 1986; McDougall et al. 1998; Huang et al. 2002). This is because icons are easy to process (Horton 1994, 1996; Lin and Yeh 2010; McDougall et al. 1999; Wiedenbeck; 1999) and convenient for universal communication (Arend et al. 1987; Horton 1994, 1996; Lodding 1983; McDougall et al. 1999).

Prior research has found that attractiveness leads into better ratings of interfaces primarily due to the use of graphic elements, such as icons (Roberts et al. 2003). Icons are one main component of GUI design, and results show that attractive and appropriately designed icons increase consumer interest and interaction within online storefront interfaces, such as app stores (Burgers et al. 2016; Chen 2015; Hou and Ho 2013; Jylhä and Hamari 2019; Lin and Chen 2018; Lin and Yeh 2010; Salman et al. 2010, 2012; Shu and Lin 2014; Wang and Li 2017). While icons do not constitute a graphical user interface solitarily, an icon-based GUI is a highly common presentation in best-selling devices at present. This justifies using icons as study material for evaluating visual qualities of graphical user interface elements. Hence, VISQUAL was validated by experimenting on user interface icons.

Prior studies have introduced different methods to measure the aesthetics of graphical user interfaces during the past decades. Please refer to Table 1 for a summary list of instruments.

Metric-based instruments include multi-screen interface assessment with formulated aesthetic measures and visual techniques (Ngo et al. 2000, 2001; Vanderdonckt and Gillo 1994), semi-automated computation of user interfaces with the online tool QUESTIM (Zen and Vanderdonckt 2016) as well as predictive computation of on-screen image and typeface aesthetics (Maity et al. 2015, 2016). Survey-based instruments include a semantic differential scale measuring hedonic and pragmatic qualities of interface appeal (Hassenzahl et al. 2003) and a scale measuring perceived simplicity of user interfaces in relation to visual aesthetics (Choi and Lee 2012).

Semantic differential is a commonly used tool for measuring connotative meanings of concepts. Similar to AttrakDiff 2 (Hassenzahl et al. 2003), semantic

Table 1 Measurements for graphical user interface aesthetics

Measure	Construct	Description	Original paper
Aesthetic measures for assessing graphic screens	Multi-screen interface assessment (metric-based)	Aesthetic measures of (1) balance, (2) equilibrium, (3) symmetry, (4) sequence, (5) order, and (6) complexity	Ngo et al. (2000)
Aesthetic measures for assessing graphic screens (extended)	Multi-screen interface assessment (metric-based)	Aesthetic measures of (1) balance, (2) equilibrium, (3) symmetry, (4) sequence, (5) cohesion, (6) unity, (7) proportion, (8) simplicity, (9) density, (10) regularity, (11) economy, (12) homogeneity, and (13) rhythm	Ngo (2001)
Visual techniques for traditional and multi-media layouts	Computation of visual techniques (metric-based)	Five sets of visual techniques measuring (1) physical techniques, (2) composition techniques, (3) association and dissociation techniques, (4) ordering techniques, and (5) photographic techniques	Vanderdonckt and Gillo (1994)
Quality estimator using metrics (QUESTIM)	Computation of aesthetic user interface metrics (metric-based, online software)	Semi-automated computation of (1) balance, (2) density, (3) alignment, (4) centrality, (5) simplicity, (6) proportion, and (7) symmetry. Accessible as online software. questimapp.appspot.com	Zen and Vanderdonckt (2014, 2016)
Nonlinear regression model for aesthetic ratings of on-screen images	Predictive computation of on-screen image aesthetics (metric-based)	Aesthetic measures of 20 qualities predicting geometry-related features and image-related features	Maity et al. (2015)
Predictive aesthetic model for textual contents on interfaces	Weighted sum of multiple textual element features (metric-based)	Aesthetic measures of (1) chromatic contrast, (2) luminance contrast, (3) font size, (4) letter spacing, (5) line height, and (6) word spacing	Maity et al. (2016)

Table 1 (continued)

Measure	Construct	Description	Original paper
AttrakDiff 2	Hedonic and pragmatic evaluation of interface appeal (survey-based, online software)	Seven-point semantic differential scale of 21 items measuring (1) hedonic quality–identification, (2) hedonic quality–stimulation, and (3) pragmatic quality. Accessible as online software. attrakdiff.de/index-en.html	Hassenzahl et al. (2003)
Scale of simplicity	Simplicity perception of interfaces (survey-based)	Seven-point scale measuring six components: (1) reduction, (2) organization, (3) component complexity, (4) coordinative complexity, (5) dynamic complexity, and (6) visual aesthetics	Choi and Lee (2012)

differential scale was utilized in the development of VISQUAL. However, in addition to differences in items, AttrakDiff 2 was developed by comparing user interfaces as entities, while the validation of VISQUAL was performed via measuring visual qualities of single GUI items. This allows for the evaluation of several varying elements within an interface regardless of layout composition and context limitations. Hence, VISQUAL may be utilized to measure visual qualities of, e.g., icons and fonts in order to compose a successful graphical user interface. Furthermore, AttrakDiff 2 measures hedonic and pragmatic qualities of entire user interfaces. While an effective user interface constitutes of a plethora of factors, measures should be taken to produce appealing designs for enhanced usability (Kurosu and Kashimura 1995; Ngo et al. 2000; Salimun et al. 2010; Tractinsky 1997; Tractinsky et al. 2000) as well as sense of pleasure and trust (Cyr et al. 2006; Jordan 1998; Zen and Vanderdonck 2016). This justifies the development of an element-specific evaluation instrument for visual aesthetics, namely VISQUAL.

Inconsistent findings within the handful of instruments developed suggest that a reliable method is yet to be found. This study aims to address gaps in prior research that has attempted to measure graphical user interface aesthetics as an entity utilizing different platforms as study material, such as website layouts. To our knowledge, no measurement has yet been proposed to explore visual qualities of single GUI elements as parts of a harmonious interface. Attractive qualities of user interfaces contribute to a positive user experience (Hamborg et al. 2014), justifying our intentions to lay the groundwork with potentially far-reaching practical and theoretical implications. Therefore, we investigated what aesthetic features appear together in graphical icons measured via user perceptions. Based on these results, we developed an instrument that measures visual qualities of graphical user interface elements. First, we devised a semantic differential scale of 22 adjective pairs. We then conducted a survey-based vignette study with random participant ($n=569$) assignment to evaluate 4 icons from a total of pre-selected 68 game app icons across 4 categories (concrete, abstract, character and text) using the semantic scales. Game app icons were used for validity and comparability in the results. This garnered a total of 2276 individual icon evaluations. The large-scale quantitative data were analyzed in two ways by exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). As a result, VISQUAL was composed. The following section introduces the study design in detail.

3 Methods and data

As a foundation for this study, a semantic differential scale of 22 adjective pairs was employed to measure visual qualities of graphical user interface elements. We conducted a within-subjects vignette study with random participant ($n=569$) assignment to evaluate 4 icons from a total of pre-selected 68 game app icons across 4 categories (concrete, abstract, character and text) using the semantic scales. Game app icons were used for validity and comparability in the results. This resulted in a total of 2276 individual icon evaluations. The following describes the participants in the study.

3.1 Participants

A nonprobability convenience sample was composed of 569 respondents who each assessed 4 game app icons through a survey-based vignette experiment. A link to the online experiment was advertised in Facebook groups and Finnish student organizations' mailing lists. The experiment was a self-administered online task. The aim was to gather data by exposing the participants close to a realistic setting outside an authentic app store context. Please refer to Table 2 for demographic details of participants.

The majority of the participants were from Finland (92.8%). Only slightly more than half of the sample body were male (52.2%) with a mean age of 26.90 years ($SD=7.24$ years; 16–62 years). Most participants were university students (61.7%) and had a university-level education (39.9%). Two participants were raffled to receive a prize (Polar Loop 2 Activity Tracker). No other participation fees were paid. Participants were informed about the purpose of the study and assured anonymity throughout the experiment.

3.2 Measure development

In order to measure visual qualities of graphical user interface elements, i.e., game app icons, a seven-point semantic differential scale was constructed (e.g., Beautiful 1 2 3 4 5 6 7 Ugly). Semantic differential is commonly used to measure connotative meanings of concepts with bipolar adjective pairs. In total, 22 adjective pairs were formulated according to the prior literature and assigned to each icon. This method was chosen on the basis of our research objective, which was to find out how much of a trait or quality an item (i.e., icon) has, and to examine how strongly these traits cluster together. The polarity of the adjective pairs was rotated so that perceivably positive and negative adjectives did not align on the same side of the scale. Prior to the analyses, items were reverse coded as necessary.

Prior research (Shaikh 2009) on onscreen typeface design and usage has introduced a semantic scale of 15 adjective pairs, which we adapted in our measurement instrument. Additionally, adjective pairs related to visual qualities of graphical user interface icons were added as suggested per the previous literature. These adjectives include concrete and abstract (Arend et al. 1987; Blankenberger and Hahn 1991; Dewar 1999; Hou and Ho 2013; Isherwood et al. 2007; McDougall and Reppa 2008; McDougall et al. 1999, 2000; Moyes and Jordan 1993; Rogers and Osborne 1987), simple and complex (Choi and Lee 2012; Goonetilleke et al. 2001; McDougall and Reppa 2008; McDougall and Reppa 2013; McDougall et al. 2016) as well as unique and ordinary (Creusen and Schoormans 2005; Creusen et al. 2010; Dewar 1999; Goonetilleke et al. 2001; Huang et al. 2002; Salman et al. 2010). Furthermore, adjective pairs that measure the aesthetics of graphical user interface elements were added. These adjective pairs include professional and unprofessional (Hassenzahl et al. 2003), colorful and colorless

Table 2 Demographic information

		<i>n</i>	%
Age (SD = 7.24) (Mean = 26.90) (Median = 25.00)	–20	60	10.54
	21–25	249	43.76
	26–30	145	25.48
	31–35	45	7.91
	36–40	37	6.50
	41–45	16	2.81
	46–50	7	1.23
	51–55	5	0.88
	56–60	3	0.53
	60–	2	0.35
Education	Less than high school	5	.9
	High school	135	23.7
	College	95	16.7
	Bachelor's degree	227	39.9
	Master's degree	98	17.2
Employment	Higher than master's degree	9	1.6
	Working full-time	133	23.4
	Working part-time	62	10.9
	Student	351	61.7
	Unemployed	11	1.9
Gender	Retired	1	.2
	Male	297	52.2
	Female	257	45.2
Yearly income	Other	15	2.6
	Less than \$19,999	330	58.0
	\$20,000 to \$39,999	105	18.5
	\$40,000 to \$59,999	57	10.0
	\$60,000 to \$79,999	25	4.4
	\$80,000 to \$99,999	13	2.3
	\$100,000 to \$119,999	14	2.5
\$120,000 to \$139,999	10	1.8	
\$140,000 or more	15	2.6	

(Allen and Matheson 1977), realistic and unrealistic as well as two-dimensional and three-dimensional (Vanderdonckt and Gillo 1994).

Table 3 lists the adjective pairs used in the study in alphabetical order as well as their sources, and presents an overview of the means and standard deviations. There were no critical outlier values, and the range between the lowest and highest scores clusters closely to the average even though the 68 icons were quite different from each other. All the mean scores are between 3.5 and 4.5 for each evaluation. Furthermore, we tested for skewness and the range between the lowest

Table 3 Adjective pairs, means and standard deviations (values were comprised between 1 and 7)

Adjective pairs	References	Mean	SD
Beautiful–Ugly	Shaikh (2009)	4.57	1.618
Calm–Exciting	Shaikh (2009)	3.96	1.452
Colorful–Colorless	Allen and Matheson (1977)	3.77	1.810
Complex–Simple	Choi and Lee (2012), Grootenilleke et al. (2001), McDougall and Reppa (2008, 2013), McDougall et al. (2016)	4.69	1.669
Concrete–Abstract	Arend et al. (1987), Blankenberger and Hahn (1991), Dewar (1999), Hou and Ho (2013), Isherwood et al. (2007), McDougall and Reppa (2008), McDougall et al. (1999, 2000), Moyes and Jordan (1993), Rogers and Osborne (1987)	4.02	1.998
Delicate–Rugged	Shaikh (2009)	4.42	1.368
Expensive–Cheap	Shaikh (2009)	4.83	1.563
Feminine–Masculine	Shaikh (2009)	4.34	1.388
Good–Bad	Shaikh (2009)	4.34	1.641
Happy–Sad	Shaikh (2009)	3.80	1.507
Old–Young	Shaikh (2009)	3.98	1.611
Ordinary–Unique	Creusen and Schoormans (2005), Creusen et al. (2010), Dewar (1999), Grootenilleke et al. (2001), Huang et al. (2002), Salman et al. (2010)	3.39	1.651
Passive–Active	Shaikh (2009)	3.97	1.708
Professional–Unprofessional	Hassenzahl et al. (2003)	4.22	1.736
Quiet–Loud	Shaikh (2009)	4.12	1.601
Realistic–Unrealistic	Vanderdonckt and Gillo (1994)	4.22	1.592
Relaxed–Stiff	Shaikh (2009)	4.47	1.560
Slow–Fast	Shaikh (2009)	3.87	1.576
Soft–Hard	Shaikh (2009)	4.19	1.545
Strong–Weak	Shaikh (2009)	3.93	1.464
Three-dimensional–Two-dimensional	Vanderdonckt and Gillo (1994)	4.67	1.863
Warm–Cool	Shaikh (2009)	4.02	1.435

and highest scores are between -0.5 and 0.5 , which indicates that the data are fairly symmetrical.

3.3 Materials

A total of 68 game app icons from Google Play Store were selected for the experiment. Four icons corresponding to common icon styles (concrete, abstract, character and text) were selected from each of the 17 categories for game apps (action, adventure, arcade, board, card, casino, casual, educational, music, puzzle, racing, role playing, simulation, sports, strategy, trivia and word). The design of graphical user interface elements is dependent on context (Shu and Lin 2014). Hence, we considered it justified to include icons from all categories in order to avoid systematic bias. Moreover, as the prior literature has highlighted the relevance of concreteness and abstractness as well as whether an icon includes face-like elements or letters, we ensured that one icon from each category was characteristic of one of these attributes. Please refer to Table 4 for the icons used in the study.

Additional criteria were the publishing date of the apps and the number of installs and reviews they had received at the time of selection. Since the icons in the experiment were chosen during December 2016, the acceptable publishing date for the apps was determined to range from December 3–17, 2016. No more than 500 installs and 30 reviews were permitted. The aim of this was to choose new app icons to eliminate the chance of app and icon familiarity and thus, systematic bias. Moreover, the goal was to have a varied sample of icons both in terms of visual styles and quality, meaning that several different computer graphic techniques were included, such as 2D and 3D rendered images.

3.4 Procedure

The data were collected through a survey-based vignette experiment. Respondents were provided the purpose of the study after which they were guided to fill out the survey. The survey consisted of three or four parts depending on the choice of response. The first part mapped out mobile game and smartphone usage with the following questions: “Do you like to play mobile games?”, “In an average day, how much time do you spend playing mobile games?” and “How many smartphones are you currently using?”. The second part included more specific questions about the aforementioned, e.g., the operating system of the smartphone(s) in use, the average number of times browsing app stores per week and the amount of money spent on app stores during the past year, as well as the importance of icon aesthetics when interacting with app icons. If the respondent answered that they do not use a smartphone in the first part, they were assigned directly to the third part.

In the third part, the respondent evaluated app icons using semantic differential scales. Prior to this, the following instructions were given on how to evaluate the icons: “In the following section you are shown pictures of four (4) mobile game icons. The pictures are shown one by one. Please evaluate the appearance of each icon according to the adjective pairs shown below the icon. In each adjective pair,

the closer you choose to the left or right adjective, the better you think it fits to the adjective. If you choose the middle space, you think both adjectives fit equally well.” The respondent was reminded that there are no right or wrong answers and was then instructed to click “Next” to begin. The respondent was shown one icon at a time and was asked to rate the 22 adjective pairs under the icon graphic with the following text: “In my opinion, this icon is...”. Each respondent was randomly assigned four icons to evaluate, one from each category of pre-selected icon attributes (abstract, concrete, character and text). After the semantic scales, the participant rated their willingness to click the icon as well as download and purchase the imagined app that the icon belongs to, by using a seven-point Likert scale on the same page with the icon. Lastly, demographic information (age, gender, etc.) was asked. The survey took about 10 min to complete. The survey was implemented via SurveyGizmo, an online survey tool. All content was in English. The data were analyzed with IBM SPSS Statistics and Amos version 24 as well as Microsoft Office Excel 2016.

4 Stage 1: Evaluating the instrument





































































The instrument was evaluated with three stages of consecutive analyses. First, we examined factor loadings of the 22 visual qualities with exploratory factor analysis (EFA) to examine underlying latent constructs (Table 5). Second, we performed a confirmatory factor analysis (CFA) with structural equation modeling (SEM) to assess whether the psychometric properties of the instrument (Fig. 1) are applicable to similar latent constructs, which revealed the need for modification in the model. Following the adjustments, another CFA was performed in order to finalize the model (Fig. 2).

Initially, the factorability of the 22 adjective pairs was examined. The data set was determined suitable for this purpose as the correlation matrix showed coefficients above .3 between most items with their respective predicted dimension. Moreover, the Kaiser–Meyer–Olkin measure of sampling adequacy indicated that the strength of the relationships among variables was high ($KMO = .87$), and Bartlett’s test of sphericity was significant ($\chi^2(231) = 21,919.22; p < .001$).

Given these overall indicators, EFA with varimax rotation was performed to explore factor structures of the 22 adjective pairs used in the experiment, using data from 2276 icon evaluations. There were no initial expectations regarding the number of factors. Principal component analysis (PCA) was used as extraction method to maximize the variance extracted. Varimax rotation with Kaiser normalization was used. Please refer to Table 5 for the results of the analysis.

The analysis exposed five distinguishable factors: Excellence/Inferiority, Graciousness/Harshness, Idleness/Liveliness, Normalness/Bizarreness and Complexity/Simplicity. Typically, at least two variables must load on a factor so that it can be given a meaningful interpretation (Henson and Roberts 2006). Correlations starting from .4 can be considered credible in that the correlations are of moderate strength or higher (Evans 1996). In this light, all the factors formed in the analysis are valid.

Table 4 Icons in the study

Category	Concrete	Abstract	Character	Text
Action				
Adventure				
Arcade				 Dropper
Board				
Card				
Casino				
Casual				
Educational				
Music				
Puzzle				
Racing				
Role Playing				
Simulation				
Sports				
Strategy				
Trivia				
Word				

Five adjective pairs (*good–bad*, *professional–unprofessional*, *beautiful–ugly*, *expensive–cheap* and *strong–weak*) loaded on the first factor. This factor was named *Excellence/Inferiority*. Seven adjective pairs (*hard–soft*, *relaxed–stiff*, *feminine–masculine*, *delicate–rugged*, *happy–sad*, *colorful–colorless* and *cool–warm*) loaded on the second factor. This factor was named *Graciousness/Harshness*. Five adjective pairs (*slow–fast*, *quiet–loud*, *calm–exciting*, *passive–active* and *old–young*) loaded on the third factor. This factor was named *Idleness/Liveliness*. Three adjective pairs (*concrete–abstract*, *realistic–unrealistic* and *unique–ordinary*) loaded on the fourth factor. This factor was named as *Normalness/Bizarreness*. Finally, two adjective pairs (*complex–simple* and *two-dimensional–three-dimensional*) loaded on the fifth factor. This factor was named *Complexity/Simplicity*.

5 Stage 2: Confirmatory factor analysis

In order to assess the latent psychometric properties of the instrument, confirmatory factor analysis (CFA) was performed. To accomplish this, covariance-based structural equation modeling (CB-SEM) was applied. Please refer to Fig. 1 for the model evaluated in the confirmatory factor analysis.

As per recommendation by the prior literature (Kline 2011), model fit was examined by the Chi square test (χ^2), comparative fit index (CFI), root mean square error of approximation (RMSEA), and standardized root mean square residual score (SRMR). The Chi square test shows good fit for the data if the p value is $> .05$. However, for models with sample size of more than 200 cases, the Chi square is almost always statistically significant and may not be applicable (Matsunaga 2010; Russell 2002). Generally, a CFI score of $> .95$ is considered good, whereas a score of > 0.90 is considered acceptable. RMSEA and SRMR are regarded good if the values are less than $.05$, and acceptable with values that are less $.10$.³

The initial results of the model fit indices were inadequate: $\chi^2 = 5381.664$, $DF = 199$; $\chi^2/DF = 27.044$, $p \leq .001$, CFI = $.762$, RMSEA = $.107$, and SRMR = $.1206$. These values are outside the acceptable boundaries. This is partially due to the relatively large sample size (2276 icon evaluations), as the χ^2 and p values are highly sensitive to sample size (Matsunaga 2010; Russell 2002). As such, these values will remain statistically significant and should thus be disregarded in favor of other indicators. However, the remaining values that are not as sensitive to sample size (CFI, RMSEA and SRMR) also fit poorly to the data.

Cronbach's alpha was used to assess the reliability of the scale. The prior literature suggests 0.7 as the typical cutoff level for acceptable values (Nunnally and Bernstein 1994). Alpha values for each dimension were as follows: Excellence/Inferiority ($\alpha = .879$), Graciousness/Harshness ($\alpha = .813$), Idleness/Liveliness ($\alpha = .818$), Normalness/Bizarreness ($\alpha = .460$), and Complexity/Simplicity ($\alpha = .496$). While

³ Kenny, D.A., "Measuring Model Fit," <http://davidakenny.net/cm/fit.htm> (accessed November 21, 2018).

Table 5 Exploratory factor analysis with varimax rotation (loadings > .4 bolded)

	Excellence/Inferiority (Variance extracted % = 17.353)	Graciousness/Harshness (Variance extracted % = 16.434)	Idleness/Liveliness (Variance extracted % = 15.720)	Normalness/Bizarreness (Variance extracted % = 7.828)	Complexity/Simplicity (Variance extracted % = 6.163)
Good–Bad	.838	.243	–.151	.124	–.021
Professional–Unprofessional	.835	.052	–.039	.045	.055
Beautiful–Ugly	.809	.328	–.074	.079	.021
Expensive–Cheap	.806	.067	–.121	.036	.240
Strong–Weak	.664	–.348	–.269	.051	.047
Soft–Hard	–.150	.793	.040	.026	–.005
Relaxed–Stiff	.203	.777	–.027	.046	.000
Feminine–Masculine	.008	.713	.192	–.098	.189
Delicate–Rugged	.310	.652	.130	–.072	.116
Happy–Sad	.296	.618	–.332	.135	–.099
Colorful–Colorless	.128	.568	–.460	.079	.164
Warm–Cool	–.075	.480	–.368	.103	–.068
Slow–Fast	–.191	.025	.811	–.064	–.056
Quiet–Loud	.096	.110	.805	–.027	–.065
Calm–Exciting	–.141	.013	.792	–.006	–.106
Passive–Active	–.214	–.138	.767	–.107	–.158
Old–Young	–.232	–.384	.419	.171	–.096
Concrete–Abstract	.000	.061	–.179	.810	.066
Realistic–Unrealistic	.242	–.019	.087	.738	.034
Ordinary–Unique	–.393	–.134	.031	.413	–.379
Complex–Simple	.101	.053	–.212	.024	.834
Three–Two-dimensional	.125	.127	–.213	.474	.552

three of the factors showed good level of internal consistency, two were found to have unacceptable alpha values.

Additionally, there were some concerns related to convergent validity where the average variance extracted (AVE) was less than .5, namely Graciousness/Harshness (AVE=.393) and Complexity/Simplicity (AVE=.361). Additionally, concerns related to composite reliability were discovered where the CR was less than .7, namely Normalness/Bizarreness (CR=.686) and Complexity/Simplicity (CR=.520). In terms of discriminant validity, the square root of the average variance extracted of each construct is larger than any correlation between the same construct and all the other constructs (Fornell and Larcker 1981). Please refer to Table 6 for full validity and reliability scores.

According to these results, two factors out of five proved to be robust, namely Excellence/Inferiority and Idleness/Liveliness. At this stage, the instrument does not seem to be an optimally fitting measurement model due to the poor model fit indices and the noted problems with validity and reliability. Additional issue here is the unacceptable loadings (Fig. 1). While loadings should fall between .32 and 1.00 (Matsunaga 2010; Tabachnick and Fidell 2007), the model contains values that are outside of these boundaries. These observations suggest for post hoc adjustments in the model.

As noted by the prior literature (Brown 2015; MacKenzie et al. 2011), the removal of poorly behaved reflective indicators may offer to improve the overall model fit. Furthermore, examining strong modification indices (MI=3.84) and covarying items accordingly (MacKenzie et al. 2011) is likely to prove beneficial in balancing unacceptable loadings in the model. By addressing issues associated with the problematic factors, low scores related to model fit as well as validity and reliability are expected to improve.

6 Stage 3: Finalizing the instrument

The confirmatory factor analysis in Stage 2 revealed a number of problems related to model fit, validity and reliability as well as item loadings. In order to address these issues, first, items that loaded poorly (under .65) onto the extracted factors were removed consecutively (Brown 2015). To retain the five-factor structure established in the EFA, item removal was not conducted on the Complexity/Simplicity factor despite the low loadings. Similarly, only one item with the lowest loading on the Normalness/Bizarreness factor was omitted. Deleted items are described in Table 7.

Second, modification indices (MI) were examined. A high value was found within the Excellence/Inferiority factor between the adjective pairs *professional-unprofessional* and *expensive-cheap*. Additionally, due to a high MI value, error terms were covaried for the adjective pairs *quiet-loud* and *calm-exciting* on the Idleness/Liveliness factor. These items were found to be semantically similar, and hence, the error terms of these items were allowed to correlate.

A confirmatory factor analysis was conducted on the finalized measure which comprised of five factors and the remaining 15 adjective pairs with two observed

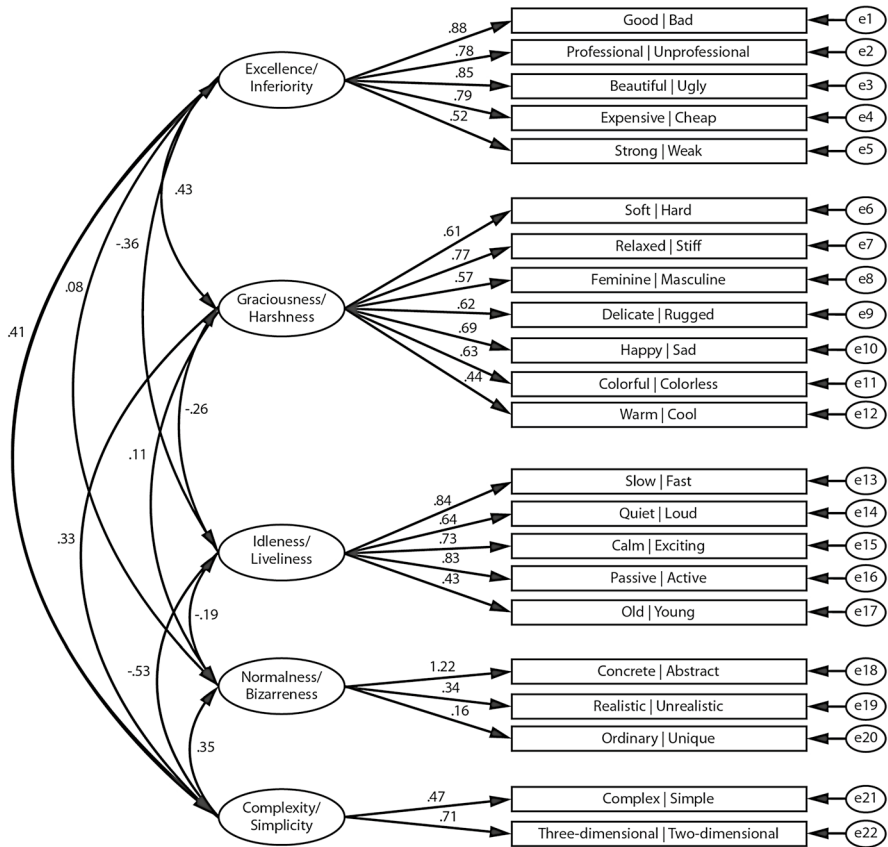


Fig. 1 Initial model with 22 items (standardized weights)

error covariances. Please refer to Fig. 2 for the adjusted model evaluated in the CFA.

With these changes, the results of the model fit indices were as follows: $\chi^2 = 1499.114$, $DF = 78$; $\chi^2/DF = 19.219$, $p \leq .001$, $CFI = .906$, $RMSEA = .089$, and $SRMR = .0705$. As discussed previously, the χ^2 and p values are highly sensitive to sample size and are thus easily inflated (Matsunaga 2010; Russell 2002). For this reason, they should be disregarded in this particular context where the instrument was assessed by using data from 2276 icon evaluations. With the exception of the discussed values, all indices showed acceptable model fit. Furthermore, all item loadings now fall between the preferred .32 and 1.00 (Matsunaga 2010; Tabachnick and Fidell 2007), although some loadings remained low (< .55) particularly on the factors with only two latent variables.

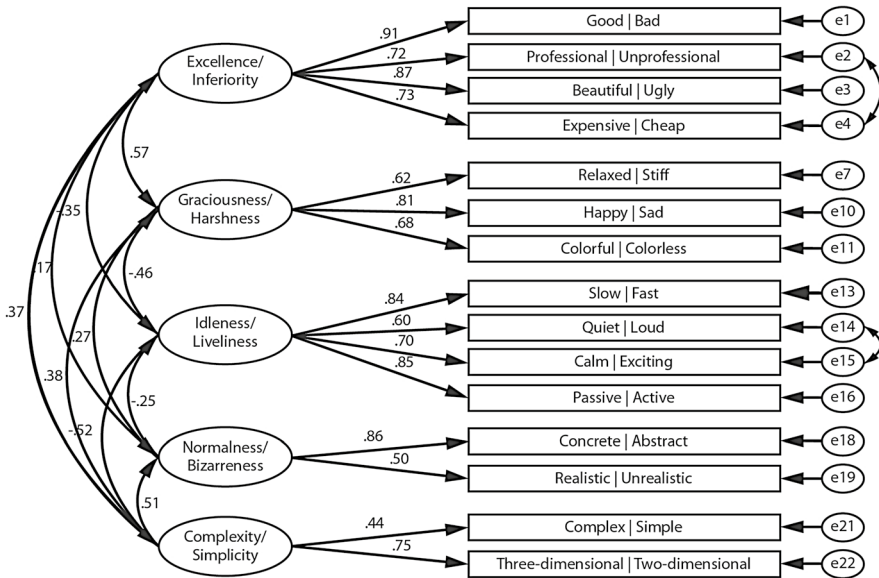


Fig. 2 Adjusted model with 15 items and covaried errors (standardized weights)

While the adjusted model retained good alpha values concerning the first three factors, previously observed issues with the last two factors remained, as follows: Excellence/Inferiority ($\alpha = .896$), Graciousness/Harshness ($\alpha = .740$), Idleness/Liveliness ($\alpha = .818$), Normalness/Bizarreness ($\alpha = .588$), and Complexity/Simplicity ($\alpha = .496$). The Complexity/Simplicity factor was not altered, thus the alpha is unchanged. However, regardless of adjustments to the model, the Normalness/Bizarreness factor did not reach an adequate alpha level.

Similarly, adjusting the model improved the AVE values, yet issues remained relating to convergent validity with three factors having AVE values under .5, namely Idleness/Liveliness (AVE = .499), Normalness/Bizarreness (AVE = .494) and Complexity/Simplicity (AVE = .378). The lower AVE score of the Normalness/Bizarreness factor in this stage is presumably caused by the removal of one semantic pair, *ordinary–unique*, which transforms the initial three-item factor into a two-item factor.

Although reliability scores showed significant increase in this stage, issues related to composite reliability remained for two factors, namely Normalness/Bizarreness (CR = .646) and Complexity/Simplicity (CR = .533). The model shows continued support for discriminant validity of the five-factor model in that the square root of AVE for each of the five factors was > 0.50 and greater than the shared variance between each of the factors. Please refer to Table 8 for full validity and reliability scores.

These results repeat the robustness of Excellence/Inferiority and Idleness/Liveliness factors. Moreover, the Graciousness/Harshness factor can be considered solid in terms of validity and reliability as the AVE value was seemingly close to the

accepted threshold of .5. Likewise, the AVE value of Normalness/Bizarreness was only slightly under the accepted threshold.

Finally, a Pearson correlation test was performed with the respondents' mean scores of both the 22-item scale and the 15-item scale to assess concurrent validity of the constructs. Please refer to Table 9 for results.

The findings show strong positive correlations between each of the 22-item constructs and their equivalents in the 15-item scale. Aside from Complexity/Simplicity ($r=1.000$, $p<0.01$) which remained unchanged throughout model adjustments, other constructs with removed items exhibit strong positive correlations as well, namely Excellence/Inferiority ($r=.982$, $p<0.01$), Graciousness/Harshness ($r=.907$, $p<0.01$), Idleness/Liveliness ($r=.969$, $p<0.01$), and Normalness/Bizarreness ($r=.894$, $p<0.01$). This observation leads to the interpretation that removal of the particular items described earlier does not critically affect the performance of the scale. Therefore, the 15-item scale can be considered as valid. While the Complexity/Simplicity factor had low loadings, it is partly accounted for by the other factors that show promise. The reason for weak loadings is presumably caused by cumulative correlation, in that Complex–Simple and Three-dimensional–Two-dimensional were perhaps perceived varyingly among the participants and poorly reflected each other, which affects the quality of the factor.

Overall, the measurement model significantly improved concerning model fit indices as well as convergent validity and composite reliability. These findings also suggest that fewer than the original number of items may be used as indicators for measuring visual qualities of graphical user interface elements. However, as there remained some concerns regarding the robustness of the finalized instrument, replication of the model with a different data sample is recommended as discussed in the following.

7 Discussion

The initial measurement model of 22 items formed a five-factor structure in the EFA in Stage 1. The factors were named to correspond to the referents on the factors: *Excellence/Inferiority*, *Graciousness/Harshness*, *Idleness/Liveliness*, *Normalness/Bizarreness* and *Complexity/Simplicity*. All items and factors were valid in the EFA. The CFA in Stage 2 exposed concerns in the model, which were countered by item removal in Stage 3. The adjusted model retained 15 (68%) items of the initial 22. As such, seven items were deleted with loadings under .65 (Table 7) on factors that held more than 2 items as the recommended solution for indicators that have low validity and reliability (MacKenzie et al. 2011). This resulted in better validity and reliability producing more robust factors, thereby theoretically justifying this choice. The majority of the removed items represent qualities that may be interpreted as ambiguous in the context of visual qualities of graphical user interfaces (e.g., *strong–weak*, *hard–soft*, *old–young*). It may be that these adjective pairs are often related to more concrete, tangible traits than visuals on an interface that are generally impalpable. Furthermore, some of these items poorly reflected others on the same factor, e.g., *strong–weak*, which can be interpreted as a synonym for quality or as a feature in a

Table 6 Validity and reliability for VISQUAL (Stage 2)

	CR	AVE	MSV	MaxR(H)	Excellence/ Inferiority	Graciousness/ Harshness	Idleness/Liveliness	Normalness/ Bizarreness	Complex- ity/Sim- plicity
Excellence/Inferiority	0.816	0.393*	0.185	0.833	0.627				
Graciousness/Harshness	0.880	0.602	0.185	0.907	0.430	0.776			
Idleness/Liveliness	0.830	0.506	0.285	0.871	-0.264	-0.358	0.711		
Normalness/Bizarreness	0.686*	0.547	0.123	1.544	0.114	0.083	-0.192	0.740	
Complexity/Simplicity	0.520*	0.361*	0.285	0.564	0.333	0.406	-0.534	0.350	0.601

*Values outside thresholds of acceptability, square root of AVE bolded

Table 7 List of deleted items, respective factors and loadings

Deleted items	Factor	Loadings
Strong–Weak	Excellence/Inferiority	.52
Warm–Cool	Graciousness/Harshness	.44
Feminine–Masculine	Graciousness/Harshness	.57
Soft–Hard	Graciousness/Harshness	.61
Delicate–Rugged	Graciousness/Harshness	.62
Old–Young	Idleness/Liveliness	.43
Ordinary –Unique	Normalness/Bizarreness	.10

visual (e.g., a character) among other explanations. Considering the other items on the factor that represent excellency in a more explicit way, this further justifies item removal from a methodological perspective.

During Stage 3, modification indices were examined for values greater than 3.84 (MacKenzie et al. 2011). Error terms were allowed to correlate between two sets of latent variables with the largest modification indices, namely *professional–unprofessional* and *expensive–cheap* as well as *quiet–loud* and *calm–exciting*. These items can be considered colloquially quite similar to their correlated pair, only that they represent similar concepts in different ways, i.e., in general and specific terms. There is an ongoing discussion whether post hoc correlations based on modification indices should be made. A key principle is that a constrained parameter should be allowed to correlate freely only with empirical, conceptual or practical justification (e.g., Brown 2015; Hermida 2015; Kaplan 1990; MacCallum 1986). Examining modification indices has been criticized, e.g., for the risk of biasing parameters in the model and their standard errors, as well as leading to incorrect interpretations on model fit and the solutions to its improvement (Brown 2015; Hermida 2015). To rationalize for these two covaried errors in the development of this particular measurement model, it is to be noted that similar to the χ^2 value and standardized residuals, modification indices are sensitive to sample size (Brown 2015). When the sample size is large (more than 200 cases), modification indices can be considered in determining re-specification (Kaplan 1990). VISQUAL was evaluated using data from 2276 icon evaluations, which causes inflation to the aforementioned values. Therefore, appropriate measures need to be taken in order to circumvent issues related to sample size. Furthermore, residuals were allowed to correlate strictly and only when the measures were administered to the same informant, i.e., factor.

This was a first-time evaluation and validation study for VISQUAL. The instrument was developed in the pursuit of aiding research and design of aesthetic interface elements, which has been lacking in the field of HCI. In this era of user-adapted interaction systems, it is crucial to advance the understanding of the relationship between interface aesthetics and user perceptions. As such, the measurement model shows promise in examining visual qualities of graphical user interface elements. However, the model fit indices were nearer to acceptable than good. In addition, convergent validity and composite reliability remain open for critique. This is perhaps an expected feature for instruments that are based on subjective perceptions

Table 8 Validity and reliability for VISQUAL (Stage 3)

	CR	AVE	MSV	MaxR(H)	Excellence/ Inferiority	Graciousness/ Harshness	Idleness/Liveliness	Normalness/ Bizarreness	Complex- ity/Sim- plicity
Excellence/Inferiority	0.747	0.499*	0.328	0.770	0.706				
Graciousness/Harshness	0.885	0.660	0.328	0.909	0.573	0.812			
Idleness/Liveliness	0.839	0.570	0.271	0.868	-0.461	-0.352	0.755		
Normalness/Bizarreness	0.646*	0.494*	0.264	0.762	0.267	0.174	-0.251	0.703	
Complexity/Simplicity	0.533*	0.378*	0.271	0.602	0.376	0.373	-0.521	0.514	0.615

*Values outside thresholds of acceptability, square root of AVE values bolded

Table 9 Pearson correlation test between 22-item scale and 15-item scale

22-item scale	15-item scale				
	Excellence/ Inferiority	Gracious- ness/Harsh- ness	Idleness/Liveliness	Normalness/ Bizarreness	Complexity/ Simplicity
Excellence/Inferiority	.982	.368	-.287	.190	.296
Graciousness/Harshness	.347	.907	-.204	.107	.242
Idleness/Liveliness	-.301	-.408	.969	-.134	-.376
Normalness/Bizarreness	.005 ^b	.046 ^a	-.088	.894	.170
Complexity/Simplicity	.295	.281	-.365	.288	1.000

All correlations statistically significant at $p < 0.01$ unless stated otherwise

^a $p < 0.05$, ^bNS

rather than more specific psychological traits. While aesthetic perception is subjective, this study shows evidence of features uniformly clustering in the evaluation of graphical user interface elements. Therefore, not only is the sentiment of what is aesthetically pleasing parallel within the responses, but also the way in which visual features in graphical items appear together. For this reason, it is advisable to observe items separately in conjunction with factors when utilizing VISQUAL in studying graphical user interface elements. Additionally, experimenting on the initial model (Fig. 1) as well as the adjusted model (Fig. 2) is recommended in further assessment of the instrument.

7.1 Implications

The growing need for customizable and adaptive interactive systems requires new ways of measuring and understanding perceptions and personality dimensions that affect how graphical user interfaces are designed and adapted. This study was one of the first attempts to develop a measurement model for individual perceptions on visual qualities of graphical user interface elements, rather than measuring an entire user interface. The scale was validated using a large sample of both graphical material (i.e., icons) and respondent data, which enhances generalizability.

Icon-based interfaces are customizable, e.g., by user navigation and theme design. Essentially, this type of user-adaptation aims for effective use, where the user-perceived pragmatic and hedonic attributes are satisfied. Features for personalization include, e.g., rearranging user interface elements per preference. Users also have the option to customize interface design by installing skins, of which data are usually gathered to determine user preferences and further recommendations on adaptation. Measured by VISQUAL, data will be available on individual perceptions of GUI elements, which can then be applied for user-adaptation. However, as modeling dynamic user preference requires both preference representation and user profile building (Liu 2015), a complementary measurement model that investigates

personality dimensions could be developed in order to strengthen our understanding on personalization.

VISQUAL is an instrument with a collaborative approach, which is frequently used in modeling individual user behavior based on group data (Zukerman and Albrecht 2001). Personality and psyche are key dimensions in user modeling and user-adaptive systems (Smith et al. 2019). As such, demographic factors as well as personality traits are to be mapped for user profiling (Chin 2001). Therefore, user perceptions derived from VISQUAL could be united with applicable methods for measuring user traits. One approach would be to combine VSQUAL with the five-factor personality model (Digman 1990) to determine personality traits for tracking user preferences of visual qualities and modifying interfaces accordingly. The five-factor model defines user personality as Openness to Experience (O), Conscientiousness (C), Extroversion (E), Agreeableness (A) and Neuroticism (N).

It has been shown that all of the five personality traits significantly affect user preferences when observing interests, e.g., those with creative tendencies (with high O) lean generally toward art and literature, whereas those with self-organized (with high C) and extroverted tendencies (with high E) lean toward health and sports (Wu et al. 2018). Demonstratively this would mean that, for example, users who are aesthetically sensitive would prefer GUI elements that are dominated by the Normalness/Bizarreness factor that highlights uniqueness, whereas users who are more self-organized and extroverted would prefer user interface elements that are dominated by the Liveliness/Idleness factor that emphasizes activity.

Therefore, the panoramic strengths of VISQUAL are threefold. First, it can be used to measure key visual elements of graphical user interfaces rather than assessing the aesthetics of an entire interface. Second, the items have been constructed in such a way that any topic of interest in GUI element design can be addressed aside from icons, e.g., menus, windows and typefaces. Finally, as the experiment is user-based, the results provide a strong overlook to user preference. This knowledge can then be adapted in establishing individual user models and designing personalized user interface systems.

This tool adds to the discourse of HCI, where usability has dominated research partly at the expense of aesthetic considerations (Hassenzahl 2004; Tractinsky et al. 2000). The development of VISQUAL has laid the groundwork for future research of evaluating graphical user interface elements and their visual qualities and how these depend on user characteristics. It may prove beneficial to scholars eager to pursue this area of work despite, or perhaps for, the need of further validating the effectiveness of this measure in different contexts of graphical user element aesthetics. A manual for administering VISQUAL is provided in “[Appendix](#)”.

7.2 Limitations and future research

VISQUAL was formulated by merging existing measures and those theorized by researchers but not previously tested, which implies limitations in the study. The initial model appeared to contain gaps that were addressed in a post hoc revision. This practice, however, moved the investigation out of a confirmatory analytic

framework. Therefore, a replication study is recommended to define the properties of the measurement model. One approach would be to split the large sample into calibration and validation samples to cross-validate the revised model (Brown 2015). This could also aid in determining the sample-dependence of modification indices and correlated errors. Although theoretically and methodologically justified, the post hoc removal of items requires further attention in exploring context-dependence. Future studies are thus recommended to utilize the model with 22 items (Fig. 1) as a means to avoid systematic bias prior to the specification of the adjusted model.

The results supported discriminant validity for the five-factor model, but concerns with convergent validity and composite reliability remain open for critique. As this was a first-time study, additional confirmatory studies are required in order to further examine the validity of the measurement model. Another subject for discussion is the overall level of reliability and validity possible to be attained by attitudinal measurement instruments where data are based on subjective intercorrelations. Intuitively, measuring user perceptions can be seen as an adequate approach for user modeling. Nevertheless, in order to strengthen our understanding on personalization, a complementary measurement model that investigates personality dimensions (i.e., attitudes, behavioral tendencies, technology acceptance, aesthetics preferences) could be developed. This would link individual user perceptions measured by VISQUAL with personality traits, which could then be used to determine further recommendations on adaptation (i.e., user modeling via stereotypes). Using VISQUAL as the basis for mapping preferential trait profiles in combination with an accurately operationalized behavior measure, it would be possible to further track the aesthetic aspects the user prefers, which can then be applied in modifying interfaces accordingly.

Additionally, VISQUAL could be revamped directly to trait measurement of preference. This would imply that, rather than asking how users perceive certain GUI elements, the instrument would measure general tendency to prefer certain qualities of GUI elements. For example, users would be asked to rate their tendencies of preference according to the five factors of VISQUAL instead of measuring the certain GUI element. This would in turn provide a preference model that could be applied on adapting GUI elements on a larger scale.

Game app icons were used in this study to maximize internal validity. This introduces a possibility for conducting future research on other app icon types for comparative results. The choice of not informing participants about the purpose of the apps behind the icons was made to avoid systematic bias. However, it would be beneficial to conduct a similar study with additional information on the app context. Finally, due to the nature and scope of this study, aesthetic measurements from other fields (e.g., website design) were not included. Other topics also important for the development of this scale that should be further assessed include demographic factors and other personal aspects such as user preferences, personality traits, and technological background. Moreover, icon understandability could be studied in order to further measure quality.

VISQUAL was validated by measuring visual qualities of single GUI elements (i.e., icons); thus, it evaluates isolated components. However, the context

surrounding the component may affect the perceived utility and usability of the component and the subjective perception of its aesthetics. As such, further research is invited to compare subjective assessments on GUI components in two scenarios: isolated and within (part of) a GUI. It is also to be studied whether the instrument is applicable in other, broader contexts as well as in other fields aside from user interface aesthetics research.

8 Conclusion

Prior research has focused on measuring graphical user interfaces as entities, although separate interface elements each have their own functions and designs. Whereas different tools and methods have been developed for assessing GUI aesthetics, no consensus exists on how to align these measures with user perceptions and the adaptation of the choice of elements to individual user preferences. The main contribution of this research is an instrument with properties that can be used to measure individual user perceptions of visual qualities—and thus, guide the design process of graphical user interface elements. However, as some concerns remained regarding validity and reliability, replication and further examination of both the initial (Fig. 1) and the adjusted model (Fig. 2) is recommended in future research.

Acknowledgements This work has been supported by Business Finland (5479/31/2017, 40111/14, 40107/14 and 40009/16) and participating partners.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix: Manual for applying VISQUAL

Please use the following reference when using, adapting, further validating or otherwise referring to VISQUAL or the paper which it was published in: Jylhä and Hamari (2020).

VISQUAL is designed for measuring perceived visual qualities of graphical user interfaces and/or singular graphical elements. The following manual guides how to apply the VISQUAL instrument. All items marked “Yes” for “Included in the final VISQUAL” should be used; however, we also recommend including the “Optional” items when administering VISQUAL. All items should preferably be presented on the same page which the graphical elements are presented on. However, if this is impractical or impossible, all measurement items should be treated equally in terms of their cognitive proximity to the graphic under investigation.

Use a seven-point semantic differential scale for each adjective pair (e.g., Beautiful 1 2 3 4 5 6 7 Ugly). The following instructions should be added beside the measured graphic: “Please evaluate the appearance of the [graphic] shown. The closer you choose to the left or right adjective, the better you think that adjective characterized the [graphic]. If you choose the middle space, you think both adjectives fit equally well.” The scale for each GUI element should be initiated with the following text: “In my opinion, this [graphic] is...”

Polarity of the adjective pairs should be randomized so that perceivably positive and negative adjectives do not align on the same side of the scale. Please refer to Table A for list of items.

Table A Items used in VISQUAL (items marked as *Optional* omitted from the adjusted model)

Factor	Adjective pair	Included in the final VISQUAL
Excellence/Inferiority	Good–Bad	Yes
	Professional–Unprofessional	Yes
	Beautiful–Ugly	Yes
	Expensive–Cheap	Yes
	Strong–Weak	Optional
Graciousness/Harshness	Soft–Hard	Optional
	Relaxed–Stiff	Yes
	Feminine–Masculine	Optional
	Delicate–Rugged	Optional
	Happy–Sad	Yes
	Colorful–Colorless	Yes
	Warm–Cool	Optional
Idleness/Liveliness	Slow–Fast	Yes
	Quiet–Loud	Yes
	Calm–Exciting	Yes
	Passive–Active	Yes
	Old–Young	Optional
Normalness/Bizarreness	Concrete–Abstract	Yes
	Realistic–Unrealistic	Yes
	Ordinary–Unique	Optional
Complexity/Simplicity	Complex–Simple	Yes
	Three-dimensional–Two-dimensional	Yes

References

- Ahmed, S.U., Mahmud, A.A., Bergaust, K.: Aesthetics in human-computer interaction: views and reviews. In: Proceedings of the 30th International Conference on HCI—New Trends in Human-Computer Interaction, San Diego, USA, pp. 559–568 (2009)
- Allen, S., Matheson, J.: Development of a semantic differential to access users' attitudes towards a batch mode information retrieval system (ERIC). *J. Am. Soc. Inf. Sci.* **28**, 268–272 (1977)
- Alvarez-Cortes, A., Zarate, V.H., Uresti, J.A.R., Zayas, B.E.: Current challenges and applications for adaptive user interfaces. In: Human-Computer interaction, Inaki Maurtua, Intech Open (2009). <https://doi.org/10.5772/7745>
- Arend, U., Muthig, K.P., Wandmacher, J.: Evidence for global superiority in menu selection by icons. *Behav. Inf. Technol.* **6**, 411–426 (1987). <https://doi.org/10.1080/01449298708901853>
- Blankenberger, S., Hahn, K.: Effects of icon design on human-computer interaction. *Int. J. Man-Mach. Stud.* **35**, 363–377 (1991). [https://doi.org/10.1016/S0020-7373\(05\)80133-6](https://doi.org/10.1016/S0020-7373(05)80133-6)
- Bouzit, S., Calvary, G., Coutaz, J., Chêne, D., Petit, E., Vanderdonck, J.: The PDA-LPA design space for user interface adaptation. In: Proceedings of the 11th International Conference on Research Challenges in Information Science (RCIS). Brighton, UK (2017). <https://doi.org/10.1109/rcis.2017.7956559>
- Brown, T.A.: *Confirmatory Factor Analysis for Applied Research*. Guilford Publications, New York (2015)
- Burgers, C., Eden, A., Jong, R., Buningh, S.: Rousing reviews and instigative images: the impact of online reviews and visual design characteristics on app downloads. *Mob. Media Commun.* **4**, 327–346 (2016). <https://doi.org/10.1177/20501579166639348>
- Chen, C.C.: User recognition and preference of app icon stylization design on the smartphone. In: Stephanidis, C. (ed.) *HCI International 2015—Posters' Extended Abstracts*. HCI 2015. Communications in Computer and Information Science, vol. 529. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-21383-5_2
- Chin, D.N.: Empirical evaluation of user models and user-adapted systems. *User Model. User-Adapt. Interact.* **11**, 181–194 (2001). <https://doi.org/10.1023/A:1011127315884>
- Choi, J.H., Lee, H.-J.: Facets of simplicity for the smartphone interface: a structural model. *Int. J. Hum. Comput Stud.* **70**, 129–142 (2012). <https://doi.org/10.1016/j.ijhcs.2011.09.002>
- Cockburn, A., Gutwin, C., Greenberg, S.: A predictive model of menu performance. In: Proceedings of the 25th Annual SIGCHI Conference on Human Factors in Computing Systems. San Jose, USA, pp. 627–636 (2007). <https://doi.org/10.1145/1240624.1240723>
- Creusen, M.E.H., Schoormans, J.P.L.: The different roles of product appearance in consumer choice. *J. Prod. Innov. Manage.* **22**, 63–81 (2005). <https://doi.org/10.1111/j.0737-6782.2005.00103.x>
- Creusen, M.E.H., Veryzer, R.W., Schoormans, J.P.L.: Product value importance and consumer preference for visual complexity and symmetry. *Eur. J. Mark.* **44**, 1437–1452 (2010). <https://doi.org/10.1108/03090561011062916>
- Cyr, D., Head, M., Ivanov, A.: Design aesthetics leading to m-loyalty in mobile commerce. *Inf. Manage.* **43**, 950–963 (2006). <https://doi.org/10.1016/j.im.2006.08.009>
- Debeve, M., Meyer, B., Donlagic, D., Svecko, R.: Design and evaluation of an adaptive icon toolbar. *User Model. User-Adap. Interact.* **6**, 1–21 (1996). <https://doi.org/10.1007/BF00126652>
- Dewar, R.: Design and evaluation of public information symbols. In: Zwaga, H.J.G., Boersema, T., Hoonhout, H.C.M. (eds.) *Visual Information for Everyday Use*, pp. 285–303. Taylor & Francis, London (1999)
- Digman, J.M.: Personality structure: emergence of the five-factor model. *Annu. Rev. Psychol.* **41**, 417–440 (1990). <https://doi.org/10.1146/annurev.ps.41.020190.002221>
- Evans, J.D.: *Straightforward Statistics for the Behavioral Sciences*. Brooks/Cole Publishing, Pacific Grove (1996)
- Fornell, C., Larcker, D.F.: Evaluating structural equation models with unobservable variables and measurement error. *J. Mark. Res.* **18**, 39–50 (1981). <https://doi.org/10.2307/3151312>
- Gait, J.: An aspect of aesthetics in human-computer communications: pretty windows. *IEEE Trans. Soft. Eng.* **8**, 714–717 (1985). <https://doi.org/10.1109/TSE.1985.232520>
- Gajos, K.Z., Crewinski, M., Tan, D.S., Weld, D.S.: Exploring the design space for adaptive graphical user interfaces. In: Proceedings of Advanced Visual Interfaces (AVI). Venezia, Italy, pp. 201–208 (2006)

- García, M., Badre, A.N., Stasko, J.T.: Development and validation of icons varying in their abstractness. *Interact. Comput.* **6**, 191–211 (1994). [https://doi.org/10.1016/0953-5438\(94\)90024-8](https://doi.org/10.1016/0953-5438(94)90024-8)
- Gittins, D.: Icon-based human–computer interaction. *Int. J. Man-Mach. Stud.* **24**, 519–543 (1986). [https://doi.org/10.1016/S0020-7373\(86\)80007-4](https://doi.org/10.1016/S0020-7373(86)80007-4)
- Goonetilleke, R.S., Shih, H.M., On, H.K., Fritsch, J.: Effects of training and representational characteristics in icon design. *Int. J. Hum. Comput. Stud.* **55**, 741–760 (2001). <https://doi.org/10.1006/ijhc.2001.0501>
- Gullà, F., Ceccacci, S., Germani, M., Cavalieri, L.: Design adaptable and adaptive user interfaces: a method to manage the information. In: Andò, B., Siciliano, P., Marletta, V., Monteriù, A. (eds.) *Ambient Assisted Living. Biosystems&Biorobotics*, vol. 11, pp. 47–58. Springer, Cham (2015)
- Hamborg, K.-C., Hülsmann, J., Kaspar, K.: The interplay between usability and aesthetics: more evidence for the “what is usable is beautiful” notion. *Adv. Hum. Comput. Int.* (2014). <https://doi.org/10.1155/2014/946239>
- Hartmann, J., Sutcliffe, A., Angeli, A.D.: Towards a theory of user judgment of aesthetics and user interface quality. *ACM Trans. Comput. Hum. Interact.* **15**, Article 15 (2007). <https://doi.org/10.1145/1460355.1460357>
- Hartmann, J., Angeli, A.D., Sutcliffe, A.: Framing the user experience: information biases on website quality judgement. In: *Proceedings of the 26th Annual SIGCHI Conference on Human Factors in Computing Systems*. Florence, Italy, pp. 855–864 (2008)
- Hassenzahl, M.: The interplay of beauty, goodness, and usability in interactive products. *Hum. Comput. Int.* (2004). https://doi.org/10.1207/s15327051hci1904_2
- Hassenzahl, M., Burmester, M., Koller, F.: AttrakDiff: EinFragebogenzurMessungwahrgenommenerhedonischer und pragmatischerQualität [AttracDiff: a questionnaire to measure perceived hedonic and pragmatic quality]. In: Ziegler, J., Szwillus, G. (eds.) *Mensch&Computer 2003*, pp. 187–196. Interaktion in Bewegung. B. G. Teubner, Stuttgart (2003)
- Henson, R.K., Roberts, J.K.: Use of exploratory factor analysis in published research: common errors and some comment on improved practice. *Educ. Psychol. Meas.* **66**, 393–416 (2006). <https://doi.org/10.1177/0013164405282485>
- Hermida, R.: The problem of allowing correlated errors in structural equation modeling: concerns and considerations. *Comput. Methods Soc. Sci.* **3**, 5–17 (2015)
- Horton, W.: *The Icon Book: Visual Symbols for Computing Systems and Documentation*. Wiley, New York (1994)
- Horton, W.: Designing icons and visual symbols. In: *Proceedings of the CHI 96 Conference on Human Factors in Computing Systems*. Vancouver, Canada, pp. 371–372 (1996). <https://doi.org/10.1145/257089.257378>
- Hou, K.-C., Ho, C.-H.: A preliminary study on aesthetic of apps icon design. In: *Proceedings of the 5th International Congress of International Association of Societies of Design Research*. Tokyo, Japan (2013)
- Huang, S.-M., Shieh, K.-K., Chi, C.-F.: Factors affecting the design of computer icons. *Int. J. Ind. Ergon.* **29**, 211–218 (2002). [https://doi.org/10.1016/S0169-8141\(01\)00064-6](https://doi.org/10.1016/S0169-8141(01)00064-6)
- Isherwood, S.J., McDougall, S.J.P., Curry, M.B.: Icon identification in context: The changing role of icon characteristics with user experience. *Hum. Fact.* **49**, 465–476 (2007). <https://doi.org/10.1518/001872007X200102>
- Jankowski, J., Bródka, P., Hamari, J.: A picture is worth a thousand words: an empirical study on the influence of content visibility on diffusion processes within a virtual world. *Behav. Inf. Technol.* **35**, 926–945 (2016)
- Jankowski, J., Hamari, J., Watrobski, J.: A gradual approach for maximising user conversion without compromising experience with high visual intensity website elements. *Int. Res.* **29**, 194–217 (2019)
- Jennings, M.: Theory and models for creating engaging and immersive ecommerce websites. In: *Proceedings of the 2000 ACM SIGCPR Conference on Computer Personnel Research*. ACM, New York, USA, pp. 77–85 (2000). <https://doi.org/10.1145/333334.333358>
- Jordan, P.W.: Human factors for pleasure in product use. *Appl. Ergon.* **29**, 25–33 (1998). [https://doi.org/10.1016/S0003-6870\(97\)00022-7](https://doi.org/10.1016/S0003-6870(97)00022-7)
- Jylhä, H., Hamari, J.: An icon that everyone wants to click: how perceived aesthetic qualities predict app icon successfulness. *Int. J. Hum. Comput. Stud.* **130**, 73–85 (2019). <https://doi.org/10.1016/j.jhcs.2019.04.004>

- Jylhä, H., Hamari, J.: Development of measurement instrument for visual qualities of graphical user interface elements (VISQUAL): a test in the context of mobile game icons. *User Model. User-Adap. Inter.* (2020). <https://doi.org/10.1007/s11257-020-09263-7>
- Kaplan, D.: Evaluating and modifying covariance structure models: a review and recommendation. *Multivar. Behav. Res.* **24**, 137–155 (1990). https://doi.org/10.1207/s15327906mbr2502_1
- Kline, R.B.: *Principles and Practice of Structural Equation Modeling*. Guilford Press, New York (2011)
- Kurosu, M., Kashimura, K.: Apparent usability vs. inherent usability. In: *Proceedings of the CHI 95 Conference Companion on Human Factors in Computing Systems*. ACM, New York, USA, pp. 292–293 (1995). <https://doi.org/10.1145/223355.223680>
- Lavie, T., Meyer, J.: Benefits and costs of adaptive user interfaces. *Int. J. Hum. Comput. Stud.* **68**, 508–524 (2010). <https://doi.org/10.1016/j.ijhcs.2010.01.004>
- Lee, S.H., Boling, E.: Screen design guidelines for motivation in interactive multimedia instruction: a survey and framework for designers. *Educ. Technol.* **39**, 19–26 (1999)
- Lin, C.-H., Chen, M.: The icon matters: how design instability affects download intention of mobile apps under prevention and promotion motivations. *Electron. Commer. Res.* (2018). <https://doi.org/10.1007/s10660-018-9297-8>
- Lin, C.-L., Yeh, J.-T.: Marketing aesthetics on the web: personal attributes and visual communication effects. In: *Proceedings of the 5th IEEE International Conference on Management of Innovation & Technology*. IEEE, Singapore, pp. 1083–1088 (2010)
- Liu, X.: Modeling users' dynamic preference for personalized recommendation. In: *Proceedings of the 24th International Joint Conference on Artificial Intelligence*. IEEE, Buenos Aires, pp. 1785–1791 (2015)
- Lodding, K.N.: Iconic interfacing. *IEEE Comput. Graph. Appl.* **3**, 11–20 (1983). <https://doi.org/10.1109/MCG.1983.262982>
- MacCallum, R.: Specification searches in covariance structure modeling. *Psychol. Bull.* **100**, 107–120 (1986). <https://doi.org/10.1037/0033-2909.100.1.107>
- MacKenzie, S.B., Podsakoff, P.M., Podsakoff, N.P.: Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *Manag. Inf. Syst.* **35**, 293–334 (2011). <https://doi.org/10.2307/23044045>
- Mahlke, S., Thüring, M.: Studying antecedents of emotional experiences in interactive contexts. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. San Jose, USA, pp. 915–918 (2007)
- Maity, R., Uttav, A., Gourav, V., Bhattacharya, S.: A non-linear regression model to predict aesthetic ratings of on-screen images. In: *Proceedings of the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction, OZCHI 2015, Parkville, Australia*, pp. 44–52 (2015). <https://doi.org/10.1145/2838739.2838743>
- Maity, R., Madrosiya, A., Bhattacharya, S.: A computational model to predict aesthetic quality of text elements of GUI. *Proc. Comput. Sci.* **84**, 152–159 (2016). <https://doi.org/10.1016/j.procs.2016.04.081>
- Matsunaga, M.: How to factor-analyze your data right: do's, don'ts, and how-to's. *Int. J. Psychol. Res.* **3**, 97–110 (2010). <https://doi.org/10.21500/20112084.854>
- McDougall, S.J.P., Reppa, I.: Why do I like it? The relationships between icon characteristics, user performance and aesthetic appeal. In: *Proceedings of the Human Factors and Ergonomics Society 52nd Annual Meeting*. New York, USA, pp. 1257–1261 (2008). <https://doi.org/10.1177/154193120805201822>
- McDougall, S.J.P., Reppa, I.: Ease of icon processing can predict icon appeal. In: *Proceedings of the 15th international conference on Human-Computer Interaction*. Las Vegas, USA, pp. 575–584 (2013). https://doi.org/10.1007/978-3-642-39232-0_62
- McDougall, S.J.P., Curry, M.B., de Bruijn, O.: Understanding what makes icons effective: how subjective ratings can inform design. In: Hanson, M. (ed.) *Contemporary Ergonomics*, pp. 285–289. Taylor & Francis, London (1998)
- McDougall, S.J.P., Curry, M.B., de Bruijn, O.: Measuring symbol and icon characteristics: norms for concreteness, complexity, meaningfulness, familiarity, and semantic distance for 239 symbols. *Behav. Res. Methods Instrum. Comput.* **31**, 487–519 (1999). <https://doi.org/10.3758/BF03200730>
- McDougall, S.J.P., de Bruijn, O., Curry, M.B.: Exploring the effects of icon characteristics on user performance: the role of icon concreteness, complexity, and distinctiveness. *J. Exp. Psychol. Appl.* **6**, 291–306 (2000). <https://doi.org/10.1037/1076-898X.6.4.291>

- McDougall, S.J.P., Reppa, I., Kulik, J., Taylor, A.: What makes icons appealing? The role of processing fluency in predicting icon appeal in different task contexts. *Appl. Ergon.* **55**, 156–172 (2016). <https://doi.org/10.1016/j.apergo.2016.02.006>
- Möttus, M., Lamas, D., Pajusalu, M., Torres, R.: The evaluation of interface aesthetics. In: Proceedings of the International Conference on Multimedia, Interaction, Design and Innovation (MIDI). Warsaw, Poland (2013). <https://doi.org/10.1145/2500342.2500345>
- Moyes, J., Jordan, P.W.: Icon design and its effect on guessability, learnability, and experienced user performance. In: Alty, J.D., Diaper, D., Gust, S. (eds.) *People and Computers VIII*, pp. 49–59. Cambridge University Society, Cambridge (1993)
- Ngo, D.C.L.: Measuring the aesthetic elements of screen designs. *Displays* **22**, 73–78 (2001). [https://doi.org/10.1016/S0141-9382\(01\)00053-1](https://doi.org/10.1016/S0141-9382(01)00053-1)
- Ngo, D.C.L., Samsudin, A., Abdullah, R.: Aesthetic measures for assessing graphic screens. *J. Inf. Sci. Eng.* **16**, 97–116 (2000)
- Ngo, D.C.L., Teo, L.S., Byrne, J.G.: Modelling interface aesthetics. *Inf. Sci.* **152**, 25–46 (2003). [https://doi.org/10.1016/S0020-0255\(02\)00404-8](https://doi.org/10.1016/S0020-0255(02)00404-8)
- Norman, D.A.: *Emotional design: why we love (or hate) everyday things*. Basic Books, New York (2004)
- Nunnally, J.C., Bernstein, I.: *Psychological Theory*. McGraw-Hill, New York (1994)
- Overby, E., Sabyasachi, M.: Physical and electronic wholesale markets: an empirical analysis of product sorting and market function. *J. Manag. Inf. Syst.* **31**, 11–46 (2014). <https://doi.org/10.2753/MIS0742-1222310202>
- Roberts, L., Rankin, L., Moore, D., Plunkett, S., Washburn, D., Wilch-Ringen, B.: Looks good to me. In: Proceedings of CHE03, Extended Abstracts on Human Factors in Computing Systems. ACM, New York, USA, pp. 818–819 (2003)
- Rogers, Y., Osborne, D.J.: Pictorial communication of abstract verbs in relation to human–computer interaction. *Br. J. Psychol.* **78**, 99–112 (1987). <https://doi.org/10.1111/j.2044-8295.1987.tb02229.x>
- Russell, D.W.: In search of underlying dimensions: the use (and abuse) of factor analysis in personality and social psychology bulletin. *Personal. Soc. Psychol. Bull.* **28**, 1629–1646 (2002). <https://doi.org/10.1177/014616702237645>
- Salimun, C., Purchase, H.C., Simmons, D., Brewster, S.: The effect of aesthetically pleasing composition on visual search performance. In: Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries. ACM, Reykjavik, Iceland, pp. 422–431 (2010). <https://doi.org/10.1145/1868914.1868963>
- Salman, Y.B., Kim, Y., Cheng, H.I.: Senior-friendly icon design for the mobile phone. In: Proceedings of the 6th International Conference on Digital Content, Multimedia Technology and its Applications (IDC 2010). IEEE, Seoul, South Korea, pp. 103–108 (2010)
- Salman, Y.B., Cheng, H.I., Patterson, P.E.: Icon and user interface design for emergency medical information systems: a case study. *Int. J. Med. Inform.* **81**, 29–35 (2012). <https://doi.org/10.1016/j.ijmedinf.2011.08.005>
- Sarsam, S.M., Al-Samarraie, H.: Towards incorporating personality into the design of an interface: a method for facilitating users' interaction with the display. *User Model. User-Adap. Interact.* **28**, 75–96 (2018). <https://doi.org/10.1007/s11257-018-9201-1>
- Schneider-Hufschmidt, M., Malinowski, U., Kuhme, T.: *Adaptive user Interfaces: Principles and Practice*. Elsevier Science Inc., New York (1993)
- Shaikh, A.D.: Know your typefaces! Semantic differential presentation of 40 onscreen typefaces. *Usab. N.* **11**, 23–65 (2009)
- Shu, W., Lin, C.-S.: Icon design and game app adoption. In: Proceedings of the 20th Americas Conference on Information Systems. Georgia, USA (2014)
- Smith, K.A., Dennis, M., Masthoff, J., Tintarev, N.: A methodology for creating and validating psychological stories for conveying and measuring psychological traits. *User Model. User-Adap. Interact.* **29**, 573–618 (2019). <https://doi.org/10.1007/s11257-019-09219-6>
- Tabachnick, B.G., Fidell, L.S.: *Using Multivariate Statistics*. Allyn and Bacon/Pearson, Boston (2007)
- Tractinsky, N.: Aesthetics and apparent usability: empirically assessing cultural and methodological issues. In: Proceedings of the ACM SIGCHI Conference on Human FACTORS in Computing Systems. ACM, New York, pp. 115–122 (1997). <https://doi.org/10.1145/258549.258626>
- Tractinsky, N., Katz, A.S., Ikar, D.: What is beautiful is usable. *Interact. Comput.* **13**, 127–145 (2000). [https://doi.org/10.1016/S0953-5438\(00\)00031-X](https://doi.org/10.1016/S0953-5438(00)00031-X)

- Vanderdonckt, J., Gillo, X.: Visual techniques for traditional and multimedia layouts. In: Proceedings of the Workshop on Advanced Visual Interfaces AVI. Bari, Italy, pp. 95–104 (1994). <https://doi.org/10.1145/192309.192334>
- Wang, M., Li, X.: Effects of the aesthetic design of icons on app downloads: evidence from an android market. *Electron. Commer. Res.* **17**, 83–102 (2017). <https://doi.org/10.1007/s10660-016-9245-4>
- Wiedenbeck, S.: The use of icons and labels in an end user application program: An empirical study of learning and retention. *Behav. Inf. Technol.* **18**, 68–82 (1999). <https://doi.org/10.1080/01449299919129>
- Wu, W., Chen, L., Zhao, Y.: Personalizing recommendation diversity based on user personality. *User Model. User-Adap. Interact.* **28**, 237–276 (2018). <https://doi.org/10.1007/s11257-018-9205-x>
- Zen, M., Vanderdonckt, J.: Towards an evaluation of graphical user interfaces aesthetics based on metrics. In: Proceedings of the IEEE 8th International Conference on Research Challenges in Information Science (RCIS). Marrakech, Morocco, pp. 1–6 (2014). <https://doi.org/10.1109/rcis.2014.6861050>
- Zen, M., Vanderdonckt, J.: Assessing user interface aesthetics based on the inter-subjectivity of judgment. In: Proceedings of the 30th International BCS Human Computer Interaction Conference. BCS, Swindon, UK (2016). <https://doi.org/10.14236/ewic/hci2016.25>
- Zukerman, I., Albrecht, D.W.: Predictive statistical models for user modeling. *User Model. User-Adap. Interact.* **11**, 5–18 (2001). <https://doi.org/10.1023/A:1011175525451>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Henrietta Jylhä is a researcher and a PhD candidate at the Gamification Group at Tampere University. Her research focuses on visual aspects in interactive environments such as graphical user interfaces relating to consumer psychology. She has experience in quantitative methods, i.e. extensive international survey studies and online experiments. She also has a degree in game and computer graphics and a strong background in digital arts. Jylhä's current research explores the relationship between consumer perceptions and app icons. <http://gamification.group/h-jylha/>.

Juho Hamari is a Professor of Gamification and leads the Gamification Group at Tampere University. He has authored several seminal academic articles on areas of gamification, games, extended realities and online economies from perspectives of human-computer interaction, information systems science, consumer behavior. His research has been published in a variety of prestigious venues such as *IEEE Transactions on Affective Computing*, *UMUAI*, *IJHCS*, *IJHCI*, *JASIST*, *IJIM*, *Organization Studies*, *New Media & Society*, *Journal of Business Research*, *Computers in Human Behavior*, *Internet Research*, *Electronic Commerce Research and Applications*, *Simulation & Gaming*, as well as in books published by among others MIT Press. <http://juhohamari.com>.

Affiliations

Henrietta Jylhä¹  · Juho Hamari¹ 

Juho Hamari
juho.hamari@tuni.fi

¹ Gamification Group, Faculty of Information Technology and Communication Sciences, Tampere University, 33014 Tampere University, Finland