

HENRIETTA JYLHÄ

Attractive User Interface Elements

Measurement and prediction

Tampere University Dissertations 758

HENRIETTA JYLHÄ

Attractive User Interface Elements
Measurement and prediction

ACADEMIC DISSERTATION

To be presented, with the permission of
the Faculty of Information Technology and Communication Sciences
of Tampere University,
for public discussion in the auditorium A1
of the Main Building, Kalevantie 4, Tampere,
on 14 April 2023, at 13.00.

ACADEMIC DISSERTATION

Tampere University, Faculty of Information Technology and Communication Sciences
Finland

<i>Responsible supervisor and Custos</i>	Professor Juho Hamari Tampere University Finland
<i>Pre-examiners</i>	Professor Meinald Thielsch University of Münster Germany Professor Isabela Gasparini Santa Catarina State University Brazil
<i>Opponent</i>	Assistant Professor Jussi Jokinen University of Jyväskylä Finland

The originality of this thesis has been checked using the Turnitin OriginalityCheck service.

Copyright ©2023 author

Cover design: Roihu Inc.

ISBN 978-952-03-2795-8 (print)

ISBN 978-952-03-2796-5 (pdf)

ISSN 2489-9860 (print)

ISSN 2490-0028 (pdf)

<http://urn.fi/URN:ISBN:978-952-03-2796-5>



Climatic CC-000255-FI
PunaMusta Printing

Carbon dioxide emissions from printing Tampere University dissertations have been compensated.

PunaMusta Oy – Yliopistopaino
Joensuu 2023

Per aspera ad astra

ACKNOWLEDGEMENTS

Advancing knowledge in new technologies and game-related environments has been my goal since the beginning of my studies. With drive and determination, this PhD is the result. This achievement, however, would not have been possible without the support of many others.

I wish to express my sincere gratitude to Tampere University for providing me with a fully funded Doctoral Researcher position for the duration of this research. Being part of the university staff presented me with experience for a lifetime.

Further, I would like to thank my supervisor, Professor Juho Hamari, for the thoughtful comments and recommendations throughout my studies. I would also like to thank the pre-examiners of this thesis, Professor Meinald Thielsch and Associate Professor Isabela Gasparini, for their valuable and insightful feedback, as well as Assistant Professor Jussi Jokinen, for agreeing to act as my opponent. I am also thankful to the Gamification Group and all its brilliant members for the considerate guidance, collaboration and warm memories.

I am lucky to have a very supportive group of friends and for that I am deeply grateful. You have all provided me with endless encouragement, inspiration and happiness. Thank you for cheering me on and for always being there for me. You guys rock!

Finally, my most heartfelt gratitude goes to my family for the enormous support throughout my life. This thesis would not have been finished without the always impeccable advice and head-on attitude from my mom, Jutta, and all the exciting adventures and countless smart discussions with my sister, Hanne-Lotte. Special thanks to the supportive group at home, Kari, and rescue cat Kaheli.

In loving memory of my grandmother, Helinä, who always made me see the bright side of life.

Helsinki, December 2022

Henrietta Jylhä

ABSTRACT

The years 2020–2021 mark a time when the global population was encountered by a world-wide pandemic. The lockdown had devastating consequences on many industries and individuals, and the emergence of global economies into the post-pandemic recovery has only just begun. However, as people adapted to the pandemic by embracing a mobile lifestyle, industries that employed graphical user interfaces as a means of human-computer interaction saw tremendous growth, exceeding everyone's expectations despite predictions of a slowdown. One example is the mobile apps and games markets, touted as the fastest growing marketplaces worldwide. At the moment, the impact of the mobile economy is undeniably high, and it does not show signs of stalling. As we look ahead and start the 'return to physical', we can see new mobile habits take shape in our everyday life.

Today, people conduct most daily functions via graphical user interfaces, due to the increasing technology-mediated nature of all human praxis, such as socializing, work, education, and entertainment. The interaction is realized on various different platforms, be they on desktop, mobile devices, VR or (smart) TVs. Although user interfaces themselves are not novel, their role is more significant now than anyone could have imagined only a few decades ago. Attractive visual designs in user interfaces have proven to enhance many aspects concerning usability, sense of pleasure and trust, but evaluating aesthetics is challenging due to the subjective nature of user perception. Although several theories and measurement instruments have been developed in order to assess and design pleasing user interfaces, the measures remain scattered. Therefore, the aim of this dissertation is to expand knowledge on *how the visual aesthetics of graphical user interfaces can be modelled, evaluated, and assessed*.

Through four studies, this dissertation provides an overview of the state-of-the-art in the literature of measurement instruments of visual aesthetics for graphical user interfaces. The dimensions of aesthetic perception that emerge in the context of user interface elements are also examined and introduced by developing a scale for measuring perceptions. As engaging and intuitive imagery has become one of the most valuable assets in today's attention economy, the studies also observe individual user perceptions of different demographic groups and their relationships on

aesthetic qualities to determine how they predict the success of graphical elements. The publications employ methodology ranging from a systematic literature review to sophisticated, quantitative statistical modelling methods to accurately identify and address each of the described phenomena by standardized means.

The findings provided by this dissertation greatly contribute to existing literature on the measurement and prediction of visually pleasing graphical user interfaces both practically and theoretically. Advancing knowledge and guidelines in this fast-paced field requires assessment from a wide perspective, including the observation of prior work, and the adaptation of measures to the modern economy by highlighting user behavior and preferences. This is particularly important in the milieu of the increasingly growing prevalence of graphical user interfaces that will continue shaping our lives in ways unimaginable.

CONTENTS

1	Introduction	15
1.1	Research problem and questions.....	17
1.2	Structure of the dissertation.....	19
2	Background	20
2.1	Visual aesthetics in graphical user interfaces.....	20
2.2	Measurement and prediction of graphical user interface aesthetics.....	23
2.3	Aspects and dimensions of the measurement instruments	28
2.4	Use contexts of the measurement instruments.....	32
2.5	Validity of the measurement instruments.....	34
2.6	Contradictions and gaps in prior research.....	37
3	Methods and data.....	39
3.1	Research approach.....	39
3.2	Literature review	40
3.3	Online survey and vignette experiment	42
3.3.1	Participants.....	43
3.3.2	Materials.....	44
3.3.3	Measurements.....	46
3.3.4	Procedure.....	48
3.4	Analyses.....	49
3.4.1	Structural equation modelling.....	49
3.4.2	Exploratory and confirmatory factor analysis	50
3.4.3	Regression analysis.....	51
4	Results	52
4.1	Publication 1	52
4.2	Publication 2.....	53
4.3	Publication 3.....	56
4.4	Publication 4.....	58
4.5	Overview of the results.....	59
5	Discussion	61
5.1	Contributions.....	63

5.2	Limitations	65
5.3	Research avenues	67
6	Conclusion and future agenda.....	69
	References	72

List of Figures

Figure 1.	Overview of publications.....	19
Figure 2.	Literature search process and outcomes	41
Figure 3.	Articles published by year.....	53

List of Tables

Table 1.	Instruments for evaluation of graphical user interface aesthetics.....	23
Table 2.	Aspects of graphical user interface aesthetics	28
Table 3.	Dimensions of UI aesthetics based on perceptions	30
Table 4.	Aspects based on deep learning and eye-tracking	31
Table 5.	Elements of graphical user interface aesthetics	32
Table 6.	Use contexts for measuring graphical user interface aesthetics	33
Table 7.	Validated and non-validated instruments.....	34
Table 8.	Demographic information.....	43
Table 9.	Icons used in the experiment.....	45
Table 10.	Constructs, means and standard deviations (adjusted items bolded).....	47
Table 11.	Items in VISQUAL	55
Table 12.	Top 6 icons with highest score (1 = lowest and 7 = highest)	57

ABBREVIATIONS

AUI	Adaptive user interface
AVE	Average variance extracted
CB-SEM	Covariance-based structural equation modelling
CFA	Confirmatory factor analysis
CFI	Comparative fit index
CR	Composite reliability
EFA	Exploratory factor analysis
GUI	Graphical user interface
HCI	Human-computer interaction
KMO	Kaiser-Meyer-Olkin measure of sampling adequacy
MI	Model fit indices
MLRA	Multiple linear regression analysis
MSV	Maximum shared variance
PCA	Principal component analysis
RMSEA	Root mean square error of approximation
SEM	Structural equation modelling
SRMR	Standardized root mean square residual score
TV	Television
UI	User interface
VIF	Variance inflation factor
VR	Virtual reality

ORIGINAL PUBLICATIONS

- Publication I Jylhä, H., and Hamari, J., 2022. Evaluation instruments for visual aesthetics of graphical user interfaces: A review. *IEEE Transactions on Visualization and Computer Graphics* (in-review).
- Publication II Jylhä, H., and Hamari, J., 2020. Development of measurement instrument for visual qualities of graphical user interface elements (VISQUAL): A test in the context of mobile game icons. *User Modeling and User-Adapted Interaction*, Vol. 30 No. 5, pp. 949–982.
- Publication III Jylhä, H., and Hamari, J., 2019. An icon that everyone wants to click: How perceived aesthetic qualities predict app icon successfulness. *International Journal of Human-Computer Studies*, Vol. 130, pp. 73–85.
- Publication IV Jylhä, H., and Hamari, J., 2021. Demographic factors have little effect on aesthetic perceptions of icons: A study of mobile game icons. *Internet Research* (ahead-of-print).

AUTHOR'S CONTRIBUTIONS

Author's contributions in the publications included in this dissertation are presented in the following according to the Contributor Roles Taxonomy (CRediT). The leading roles are marked in bold.

	Henrietta Jylhä	Juho Hamari
Conceptualization	I, II, III, IV	I, II, III, IV
Methodology	I, II, III, IV	I, II, III, IV
Formal analysis	I, II, III, IV	
Investigation	I, II, III, IV	
Original draft	I, II, III, IV	
Review & editing	I, II, III, IV	I, II, III, IV
Visualization	I, II, III, IV	
Supervision		I, II, III, IV

1 INTRODUCTION

Graphical user interfaces (GUIs) are ever more present in our everyday lives due to the increasing technology-mediated nature of all human praxis. People conduct important functions of their life such as socializing, work, education, and entertainment by interacting with graphical user interfaces, be they on desktop, mobile devices, VR or (smart) TVs. Human-computer interaction (HCI) through different types of interfaces has enabled world-wide communication that has proven essential particularly during the COVID-19 pandemic. Various factors have influenced the growth that GUI design has encountered, such as advances in computer hardware and software as well as industry and consumer preferences. Although GUIs themselves are not novel, their role is more significant than anyone could have imagined a few decades ago.

Engaging imagery has become one of the most valuable assets in today's attention economy, where companies are trying to get us hooked with targeted and tailored advertising, products, and services, all realized by the technology in our immediate reach. The modern, progressive clickbait tactics of attention marketing exist to keep people spending more time on different platforms (Timely, 2021), in which they seem to have succeeded. For example, observing what has been viewed as the fastest growing marketplaces in the world, namely app markets, shows that people spend an incredible amount of time and money on mobile platforms. Most of the revenue for the two leading app stores come from games, as they contribute approximately 60% of App Store's revenue and 80% for Google Play. The total revenue from entertainment apps is expected to rise to \$12 billion in 2022, which doubles the total for 2020 (App Annie, 2021). Considering these statistics, the impact of app and game industry to economic growth is immense.

After app stores became dominant in providing software, the number of mobile apps has been rising at a fast pace (Moreira et al., 2014). Effective design is necessary for consumer engagement, which has been noticed by online storefronts that try to attract users in different ways (Overby and Sabyasachi, 2014). Mobile application adoption has been found to be a complex entity of varying perceptions, such as gender, content price and quality, as well as time spent playing mobile games (Pappas

et al., 2019). This sets the milieu for the dissertation, as the operationalization was situated in the mobile apps and games markets, and the topic is relevant in the wider perspective of the phenomenal atmosphere that is enabled by that of games and gamification.

Aesthetics in graphical user interface design and research has quickly started to gain attention after the trend of prior literature focusing heavily on usability, perhaps at the expense of aesthetics (Tractinsky et al. 2000). The definition extends from fonts to illustrations, transforming information into visual communication through balanced (i.e., *equally weighted*), symmetrical (i.e., *equally distributed*), and appealing (i.e., *attractive*) graphics. Attractive visual design has proven to enhance e.g., usability (Kurosu and Kashimura 1995; Ngo et al. 2000; 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 Vanderdonckt 2016). However, evaluating aesthetics is challenging due to the subjective nature of user perception (Jylhä and Hamari 2019; Reinecke and Gajos 2014), and although several measurement instruments have been introduced to assess and design pleasing graphical user interfaces, the measures remain scattered. There is a need for quantification of visual aesthetics relating to GUI design (Wang et al. 2018) to advance effortless human-computer interaction that enables various daily actions for different target groups, considering people with various premises regarding age, gender, abilities and technology skills.

Graphical user interfaces with balanced elements have found to promote user engagement, while a cluttered interface may result in frustration (Jankowski et al. 2016; Jankowski et al. 2019; Lee and Boling 1999; Ngo et al. 2000; Salimun et al. 2010). Moreover, adaptive user interfaces lead into better ratings of satisfaction as well as long-term usage of platforms (Debevc et al. 1996; Hartmann et al. 2007; Sarsam and Al-Samarraie 2018). This highlights the well-established knowledge in several related fields: aesthetics matter (Hartmann et al. 2007; Tractinsky et al. 2000). For this particular reason, collaboration between artists, scientists and technologists is essential in this regard (Ahmed et al. 2009). Increasing demands for customization within HCI, marketing and interactive entertainment introduce new possibilities and challenges to scholars and practitioners.

Therefore, understanding *how the visual aesthetics of graphical user interfaces can be modelled, evaluated, and assessed* is of utmost importance, as their effects on us are also increasingly prevalent. While there have been discreet ventures to addressing these gaps in our knowledge, currently this continuum is lacking a coherent overview of the field as well as agenda for future research. This dissertation provides a thorough

approach to attractive user interfaces, their measurement and prediction through four studies that each investigate topical questions with the objective of filling major gaps in a field that is advancing at an incredibly fast pace with such possibilities in the future that might exceed the expectations of many.

1.1 Research problem and questions

In the course of the publications included in this dissertation, it became evident that visual aesthetics is indeed a broad domain that surrounds subjectivity in individual opinions, but also objectivity in general characteristics. Aesthetics is one of the important dimensions of GUI design that influences perceptions and shapes user experience.

At present, there is no standard measurement model to evaluate the visual aesthetics of GUIs. Current methods include qualitative approaches based on human responses, as well as quantitative approaches based on e.g., metric calculation and deep learning. Nevertheless, a thorough overview of the proposed measures is missing that would systematically map out the various instruments. For this reason, publication 1 aims to answer the first research question of this dissertation:

RQ1: What is the current state-of-the-art in the literature of measurement instruments of visual aesthetics for graphical user interfaces?

Measurement instruments have been proposed to assess and design pleasing 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 considering the subjective experience. In continuation of the thematic around methodology development, publication 2 complements previous works from a new perspective, by investigating what aesthetic perceptions appear together. Consequently, a new instrument VISQUAL is developed for measuring user perceptions of visual qualities of graphical user interfaces. To lay out the development process, the second research question is as follows:

RQ2: What are the psychometric properties of VISQUAL, and what dimensions of aesthetic perception emerge (in the context of GUI elements)?

In general, knowledge about the dimensions of aesthetic perception is crucial, as nowadays purchase decisions and therefore commercial success is heavily dependent on effective visuals (Overby and Sabyasachi, 2014). In connection with online storefronts (i.e., *app stores*) that have seen tremendous growth during the past years and especially during the times of the pandemic, publication 3 explored the aesthetic qualities that are likely to engage users into interacting with GUI elements (i.e., *app icons*). Drawing from the domains of user behavior and marketing, answered by publication 3, the third research question in this dissertation is:

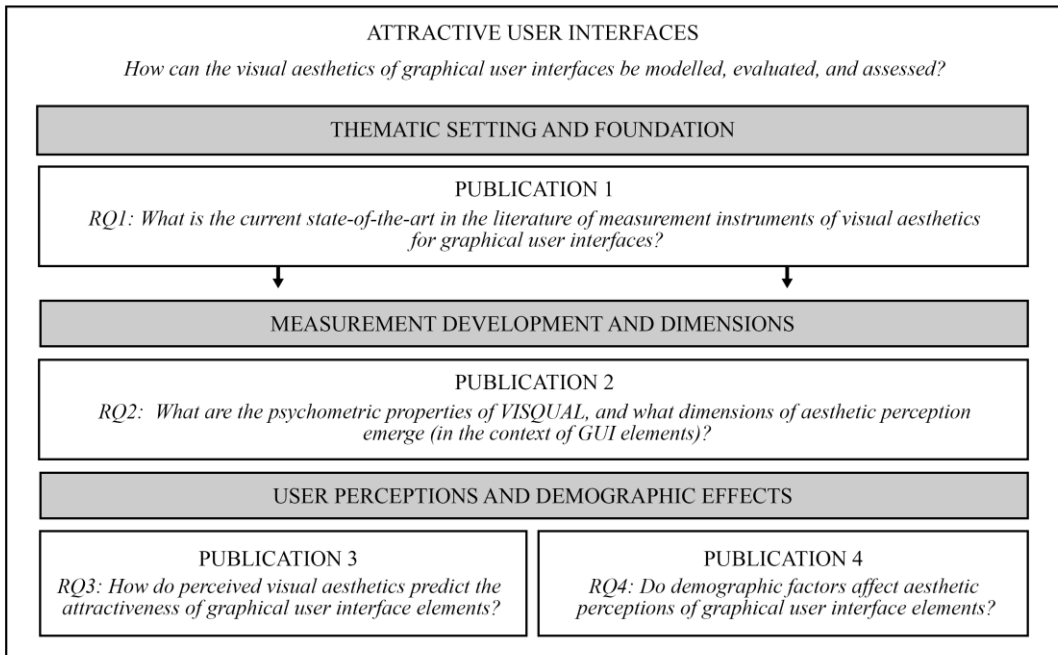
RQ3: How do perceived visual aesthetics predict the attractiveness of graphical user interface elements?

The attractiveness of GUI elements has far-reaching consequences. As such, considerations of demographic differences (i.e., age, gender and level of technology adoption) have become prevalent due to increasing demands for customization and adaptation within HCI (Norman, 2004; Tractinsky et al., 2000). However, these effects still remain relatively unexplored. This leads us to the final research question:

RQ4: Do demographic factors affect aesthetic perceptions of graphical user interface elements?

Ineffective interface usability tends to affect older age groups due to visual acuity changes (Johnson and Finn, 2017; Huang, 2013). Moreover, age is likely to contribute to users' skill level and experience with technology (KnowItAll Ninja, 2016). Therefore, age can be considered a meaningful factor in GUI aesthetics. Regarding gender differences and aesthetics in the field of HCI, preferences of male and female users has found to differ significantly previously (Genuine, 2013). However, new trends of more unisex patterns have been discovered recently (Morris et al., 2005). Time interacting with interfaces has shown to affect user attributes, preferences and expectations (Hartmann et al., 2008; Thüring and Mahlke, 2007) that can lead to different outcomes concerning interface design. As the frequency of use is related to aesthetic perceptions as well, it is an important variable in determining the subjective experience. As an overview of the dissertation, Figure 1 presents the concepts, research questions, and publications.

Figure 1. Overview of publications



1.2 Structure of the dissertation

The dissertation is divided into five chapters. The following chapter 2, *background*, introduces the wider discourse around visual qualities of graphical user interfaces as well as measurement and prediction of graphical user interface aesthetics. Chapter 3, *methods and data*, describes the literature review process and the online experiment along with details on participants, materials, measurement, procedure and analyses performed in the publications included in this dissertation. Chapter 4, *results*, presents a summary of the findings by each publication. Chapter 5, *discussion*, covers contributions of the publications, limitations, and research avenues. Chapter 6 concludes the dissertation with remarks on future research agenda.

2 BACKGROUND

The discourse of visual aesthetics in graphical user interfaces is a complex entity that combines several scientific areas, such as psychology, design and technology, as well as user experience. Modelling, evaluating, and assessing these entities further requires understanding of theories, measurement models, and context types regarding user interface aesthetics. The following sections lay out the foundation to these topics as an introduction to prior research relevant to the study. The first section defines visual aesthetics in the context of graphical user interfaces, describing user interaction from different perspectives according to prior findings. The second section discusses measurement and prediction of graphical user interface aesthetics in a comprehensive way, presenting measurement instruments by prior literature. The third and fourth section present different aesthetic aspects and dimensions assessed by the instruments, as well as use contexts of the instruments. The fifth section discusses the validity of these instruments. Finally, on the basis of the foundation laid out, the last section identifies gaps in prior research that are aimed to be filled by the publications in this dissertation.

2.1 Visual aesthetics in graphical user interfaces

Derived from art and evolutionary science, aesthetic pleasure can be defined as *the pleasure people get from processing the object for its own sake, as a source of immediate experiential pleasure in itself, and not essentially for its utility in producing something else that is either useful or pleasurable* (Dutton, 2009). Building upon this notion, aesthetic pleasure has shown to be a direct response to an object, which often precedes judgments of its utilitarian qualities or the needs it can fulfill, measured separately from an emotional or cognitive response (Blijlevens et al., 2017). The definition of visual aesthetics in the context of GUIs is *attractive computer-based environments*, reflecting the look and feel of a design, as well as the overall experience with a system (Ahmed et al., 2009; Hartmann et al., 2007b; Jennings, 2000). It is a research field that focuses on the user's subjective judgment on how aesthetically pleasing a system or a product is (Lee and Koubek, 2011), a dominant area in HCI due to the wide use of technology

for everyday actions. Aesthetics within human-computer interaction is usually divided into classical aesthetics and expressive aesthetics (Ahmed et al., 2009; Hartmann et al., 2008; Lavie and Tractinsky, 2004). Classical aesthetics refers to clear, traditional designs, whereas expressive aesthetics refer to more creative, abstract designs.

According to an information-processing model, aesthetic experiences involve five stages: perception (i.e., complexity, contrast, symmetry, order and grouping), explicit classification (i.e., style, content), implicit classification (i.e., familiarity, fluency, prototypicality), cognitive mastering (art-specific interpretation, self-related interpretation) and evaluation (understanding, ambiguity, satisfaction, pleasure) (Leder et al., 2004). The aspects of these five stages are predominant in user interface research, where e.g., symmetry and complexity are among the most studied elements (see section 2.3). These stages are further divided into aesthetic emotion and aesthetic judgments, which explains social interaction discourse and how, for example, personal taste is developed in art. Aesthetic emotion depends on the success of the information processing and can result in positive or negative feelings depending on the processing. The process can be considered rewarding, but the result can be negative in the case of aesthetic judgment. Interaction with user interfaces is done via graphical elements providing intuitive and immediate visual feedback, such as windows, menus and icons (Linux Information Project, 2004). Especially concerning interface icons, which were used as study material in publications 2, 3 and 4, attractiveness has been described as a mild aesthetic experience that refers to the power to attract users (McDougall et al., 2016). Icons are pictographic symbols, usually seen in graphics-based interfaces of operating systems (Gittins, 1986). Icons are popular in human-computer interaction, and they have replaced commands and menus as the means by which the computer supports a dialogue with the end-user (García et al., 1994; Gittins 1986; McDougall et al., 1998; Huang et al., 2002). The reason why icons are in such wide use is because they facilitate human-computer interaction being easily recognized and memorized (Horton, 1994; 1996; McDougall et al., 1999; Wiedenbeck, 1999). Icons are also convenient for universal communication, since there is no language barrier (Arend et al., 1987; Horton, 1994; 1996; Lodding, 1983; McDougall et al., 1999). Icons are one of the main elements of GUI design (Hou and Ho, 2013; Jylhä and Hamari, 2019; Shu and Lin, 2014), and results show that attractive icons increase consumer interest (Burgers et al., 2016; Chen, 2015; Wang and Li, 2017) and interaction within GUIs (Lin and Chen, 2018; Lin and Yeh, 2010; Salman et al., 2010; Salman et al., 2012). 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, which further justifies the use of icons as study material.

In system design, information structure has been connected with perceived aesthetics and usability (Ahmed et al., 2009; Cyr, 2009), and aesthetics in GUI design has been proven an integral part of a positive user experience (Kurosu and Kashimura, 1995; Ngo et al., 2000; Overby and Sabyasachi, 2014; Salimun et al., 2010; Tractinsky, 2000). Positive user experience is important for successful human-computer interaction, because encountering a negative experience may result in user frustration and abandonment of the interface. User experience is linked with visual aesthetics to an increasing extent (Debevc et al., 1996; Hartmann et al., 2007a; Sarsam and Al-Samarraie, 2018), hence, an attractive user interface is important when aiming for successful human-computer interaction as well as commercial performance (Gait, 1985; Lin and Yeh, 2010).

Users tend to carefully consider the presentation of products, and they often form their opinions on brands based on the look and feel (Orth and Malkewitz, 2008). Aesthetics is a major driver in product selection and purchase decisions (Ares et al., 2011; Creusen and Schoormans, 2005; Creusen et al., 2010; Fenko et al., 2010; Orth and Malkewitz, 2009; Schifferstein et al., 2013; van Rompay et al., 2009). This is why the appeal of GUIs should be of great importance to designers and developers. In theory, product presentation can be divided into *visual* and *informational* elements. Visual elements include e.g., layout, color, typography, size and shape, whereas informational elements include written information about the product (Silayoi and Speece, 2004). Effective visual elements in product presentation evoke more of an emotional response than informational elements (Silayoi and Speece, 2004), for example, users perceive highly saturated colors as exciting (Labrecque and Milne, 2012), making them popular in product presentation. This in turn brings extra value to the product and increases the possibility of purchase (Cho and Lee, 2005).

Regarding interface design and different users, research has shown that younger people have a more critical outlook towards aesthetics than older people. Thus, interface designers should put effort in aesthetics considerations in order to appeal to younger audiences (Oyibo et al., 2018). Moreover, older people have proven to experience more anxiety relating to human-computer interaction than younger people, therefore a number of design guidelines has been proposed (e.g., the use of large fonts, maintaining visual consistency) in order to accommodate the aging population (Johnson and Finn, 2017). Relating to interface design and gender, males have found to prefer usability, while females prefer aspects concerning beauty (Creusen, 2010; Henry, 2002; Oyibo and Vassileva, 2017; Tuch et al., 2010;

Wallendorf and Arnould, 1988). Moreover, females have been found to be more sensitive to color and visual complexity in the context of user interfaces than males (Creusen, 2010; Reinecke and Gajos, 2014; Smith, 1995).

The more the users spend time with devices has shown to affect user preferences and expectations of visual aesthetics (Lee and Koubek, 2011). Users have also been found to be more selective with aesthetics based on experience (Hartmann et al., 2008). It can be noted that involvement with GUI elements may impact users in several ways in regard to skill level, user experience, decision-making processes and perceptions of aesthetics. However, this topic has received relatively little attention especially considering mobile interfaces (Miniukovich and De Angeli, 2014b).

2.2 Measurement and prediction of graphical user interface aesthetics

Graphical user interface design has experienced change during the past decades, yet principles of visual aesthetics are still lacking. There is a need for an epistemological corpus on human factors and the quantification of aesthetic aspects relating to GUI design (Wang et al., 2018). Although several studies assessing the attractiveness of specific GUIs or GUI elements exist, instruments for the evaluation of graphical user interface aesthetics are scarce and scattered. Table 1 summarizes the instruments introduced by prior literature. The instruments are divided into evaluation based on graphical features (i.e., metrics) and evaluation based on human perceptions. Evaluation that is based on graphical features avoids human involvement in the assessment, whereas evaluation based on perceptions draw conclusions based on human responses. Additionally, evaluations conducted via emerging technologies that combine the aforementioned methods and/or introduce another approach were labelled as other (i.e., deep learning and eye-tracking).

Table 1. Instruments for evaluation of graphical user interface aesthetics

Instrument type	Studies
Evaluation based on graphical features	Bessghaier et al. (2021); Mbenza and Burny (2020); Maity et al. (2015); Maity et al. (2016); Maity and Bhattacharya (2019); Meier (1988); Miniukovich and De Angeli (2014a); Miniukovich and De Angeli (2014b); Miniukovich and De Angeli (2015a); Miniukovich and De Angeli (2015b); Miniukovich et al. (2018); Ngo et al. (2000); Ngo and Byrne (2001); Purchase et al. (2011); Reinecke et al. (2013); Reinecke and Gajos (2014); Riegler and Holzmann (2018); Uribe et al. (2017); Zen and Vanderdonck (2014); Zen and Vanderdonck (2016)

Evaluation based on human perceptions	Hassenzahl et al. (2003); Hassenzahl (2004); Jylhä and Hamari (2020); Lavie and Tractinsky (2004); Moshagen and Thielsch (2010); Moshagen and Thielsch (2013); Park et al. (2004); Sutcliffe (2002)
Other (i.e., deep learning, eye-tracking)	Dou et al. (2019); Gu et al. (2020); Khani et al. (2016); Liu and Yiang (2021); Pappas et al. (2020); Wu et al. (2011); Wu et al. (2016); Xing et al. (2021)

There have been many attempts to measure aesthetics of graphical user interfaces by several geometry-related and image-related metrics, e.g., balance, equilibrium, symmetry to avoid human involvement in the process. These are evaluations based on graphical features. A user interface is said to be in a state of repose when all of these metrics are configured accurately. If these metrics are not perfected, it will result in a state of chaos (Ngo et al., 2000). The first study concerning the measurement of GUI aesthetics was a metric-based evaluation by Meier (1988) named ACE, which applies color theory to user interface design via automation. The prototype used pre-programmed color rules as constraints to determine the best colors for user interface elements. In a more modern setting, one of the earliest studies to explore metric-based aesthetic assessment was the work of Ngo et al. (2000) and Ngo and Byrne (2001). These studies presented a subset of 14 metrics to quantify different layout aspects of GUIs. The metrics were then calculated based on e.g., the distance from the central line of the GUI or the number of layout elements. Following this work, several other combinations of the metric aspects have been introduced and investigated, usually considering at least ten metrics to calculate visual aesthetics (Bessghaier et al., 2021; Maity et al., 2015; Mbenza and Burny, 2020; Purchase et al., 2011; Zen and Vanderdonck, 2014; Zen and Vanderdonck, 2016). Relating to mobile user interface aesthetics, Bessghaier et al., (2021) developed the Aesthetic Defects DEtection Tool (ADDEI) to determine the structural aesthetic dimension automatically. The tool includes combined metric assessment to check various structural properties. Maity et al. (2015) performed an image-based evaluation with several metrics and color-related features calculated in MatLab. Mbenza and Burny (2020) combined 10 metrics creating AesthetiXML web service, while Purchase et al. (2011) calculated an overall score of 14 combined metrics to assess GUI aesthetics via a web-based script. Zen and Vanderdonck (2014; 2016) developed a web application named the Quality Estimator Using Metrics (QUESTIM) that computes regions and metrics of desktop and mobile GUIs. Miniukovich and De Angeli (2014a) proposed another metric system based on psychological assessment of visual complexity in the creation of tLight, an automatic GUI evaluation tool. The formula described metric aspects belonging to three dimensions: information amount (visual clutter and color variability), information

organization (symmetry, grid, ease-of-grouping and prototypicality), and information discriminability (contour density and figure-ground contrast). The work was further extended and tested to cover both web and mobile graphical user interfaces (Miniukovich and De Angeli, 2014b; Miniukovich and De Angeli, 2015a; Miniukovich and De Angeli, 2015b) as well as operationalized into another subset of complexity measures (Miniukovich et al., 2018). Visual complexity and colorfulness were also central elements in other measures (Reinecke et al., 2013; Reinecke and Gajos, 2014; Riegler and Holzmann, 2018, Uribe et al., 2017) demonstrating that predictions of colorfulness and complexity can account for nearly half of the variance in observed ratings of visual appeal (Reinecke et al., 2013). Metric-based instruments that focus on the visual aesthetics assessment of specific elements (e.g., text, images, white space, color) rather than evaluating the entire GUI include the works of Maity et al. (2016) and Maity and Bhattacharya (2019). These studies calculate metrics such as typography character density, word and letter spacing, as well as font size (Maity et al., 2016; Maity and Bhattacharya, 2019).

In addition to metric-based instruments, aesthetic value of graphical user interfaces has been measured by survey-based approaches based on human perceptions. A seven-point semantic differential scale AttrakDiff 2 was introduced by Hassenzahl et al. (2003) with 21 items measuring hedonic quality–identification, hedonic quality–stimulation, and pragmatic quality. The instrument was developed further by Hassenzahl (2004) with a version that included two evaluational constructs (ugly–beautiful and bad–good), resulting in 23 semantic differential items. The research investigated graphical user interfaces of MP3 software and found that beauty is related to hedonic qualities rather than pragmatic qualities (Hassenzahl, 2004). Semantic differential is a commonly used tool for measuring connotative meanings of concepts. Similar to AttrakDiff 2 (Hassenzahl et al., 2003), semantic differential scale was used in the development of the five-dimensional scale VISQUAL (Jylhä and Hamari, 2020), the measurement instrument developed in publication 2. Initially, VISQUAL had 22 items but was further validated with 15 items. Despite AttrakDiff 2 and VISQUAL both being built on semantic differential, 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.

A survey-based method that has been widely used was developed by Lavie and Tractinsky (2004) with 25 items measuring aesthetic value of website GUIs. The participants evaluated the design of two websites based on a 7-point Likert scale ranging from “strongly disagree” to “strongly agree”. This research divided visual aesthetics into classical and expressive aesthetics, where classical aesthetics refers to orderly and clear designs and expressive aesthetics refers to creative and original designs. Similarly, a 7-point Likert scale questionnaire was composed by Moshagen and Thielsch (2010), namely The Visual Aesthetics of Website Inventory (VisAWI) questionnaire. This scale measures visual aesthetics via four dimensions: simplicity, diversity, color and craftsmanship. The survey contains 18 items to which respondents assessing websites indicate their level of agreement. A short version of this questionnaire (VisAWI-s) was later proposed by Moshagen and Thielsch (2013) with four items, one for each dimension, that provide a brief visual aesthetics assessment of websites. Park et al. (2004) also employed a 7-point Likert scale in their questionnaire, which consists of 13 aesthetic dimensions with 30 items to evaluate visual aesthetics. The items were constructed with professional designers and further validated by a survey study. Finally, a 5-point Likert scale questionnaire was developed by Sutcliffe (2002), proposing heuristics for assessing the attractiveness of user interfaces based on aesthetic design. The seven aesthetic design qualities include balanced use of colour; symmetry and style; structured and consistent layout; depth of field; choice of media to attract attention; use of personality in media to attract and persuade; and design of unusual or challenging images that stimulate the users’ imagination and increase attraction.

Recently, other methods have started to appear alongside the two main categories. These alternative methods apply e.g., deep learning (Dou et al., 2019; Khani et al., 2016; Liu and Yiang, 2021; Wu et al., 2011; 2016; Xing et al., 2021) and eye-tracking (Gu et al., 2020; Pappas et al., 2020). Dou et al. (2019) adopt a deep neural network protocol named Webthetics, using backpropagation as learning algorithm. This convolutional neural network (CNN) architecture is trained from user rating data, extracting representative features from webpages to quantify their aesthetics. Likewise, Khani et al. (2016) use a similar deep learning technique to

assess website aesthetics classifying the represented data using a Support Vector Machine (SVM) and Gaussian radial basis function kernel algorithm. Liu and Yiang (2021) divide aesthetic assessment into multiple modalities (i.e., text content scoring, image aesthetic assessment and video quality assessment) before using SVM to assess results in their deep learning model. Wu et al. (2016) also divide aesthetics into multimodal features (i.e., structural, local and global visual, and functional features) and analyze multiuser ratings with structural SVM. Additionally, the number and aspect ratio of text blocks has been applied to analyze their influence on GUI aesthetics (Wu et al., 2011). Lastly, Xing et al. (2021) presented a model using CNN to quantify user perceptions of layout, color, and texture. Several frameworks were compared in the study, and the optimal result was achieved by SE-VGG19. Aside from deep learning methods, Gu et al. (2020) use eye-tracking data, i.e., visual attention entropy (VAE), to measure the interest of users in websites via gaze points and eye movement speed. They calculate a relative visual attention entropy value (rVAE) through a heatmap to correlate with human assessments. Pappas et al. (2020) track gaze behavior of users (e.g., pupil diameter, fixation, saccade) while looking at high, neutral, and low visually appealing websites. Using Random Forest regression algorithm, the collected data is assessed according to the VisAWI questionnaire (Moshagen and Thielsch, 2010) in terms of simplicity, diversity, colorfulness, and craftsmanship. Measurement instruments should be selected based on the purpose of the research. It has been discussed that automated measures are preferable when the research objective is related to the user's ability to encode new information, and human evaluation is preferable upon finding out motivations for user behavior (Selnes and Grønhaug 1986). Both measurement types have their strengths and weaknesses. VISQUAL is essentially an instrument that measures qualities based on human perceptions, thus it vastly differs from other measurement instruments that observe the design of graphical features while avoiding human involvement in the process. Measuring with VISQUAL is therefore fundamentally reliant on respondent data, which has both advantages (e.g., acquiring user interpretation intel, gaining a deeper understanding of psychological aspects) and disadvantages (e.g., low and selective participation rate that may distort estimates) as opposed to other, more automated methods. Here we come to the conclusion that in the ideal setting, measures should be combined to attain a higher level of understanding of the research topic, which has been attempted by the aforementioned emerging technologies (i.e., deep learning and eye-tracking).

2.3 Aspects and dimensions of the measurement instruments

Tables 2, 3 and 4 present the different aesthetic aspects and dimensions assessed by the instruments presented and discussed in the previous section. They were divided according to the instrument types described in Table 1, i.e., evaluation based on graphical features (layout, imagery and geometry as well as typography), evaluation based on human perceptions, and other (i.e., deep learning and eye-tracking).

Table 2. Aspects of graphical user interface aesthetics

Type	Aspect	Studies
Evaluation based on graphical features (e.g., layout, imagery and geometry)	Alignment	Mbenza and Burny (2020); Riegler and Holzmann (2018); Zen and Vanderdonckt (2014)
	Balance	Bessghaier et al. (2021); Maity et al. (2015); Ngo et al. (2000); Ngo and Byrne (2001); Purchase et al. (2011); Reinecke et al. (2013); Reinecke and Gajos (2014); Riegler and Holzmann (2018); Zen and Vanderdonckt (2014); Zen and Vanderdonckt (2016)
	Brightness	Meier (1988)
	Clutter	Miniukovich and De Angeli (2014a); Miniukovich and De Angeli (2014b); Miniukovich and De Angeli (2015a)
	Cohesion	Bessghaier et al. (2021); Ngo and Byrne (2001); Purchase et al. (2011)
	Color	Maity and Bhattacharya (2019); Meier (1988); Miniukovich and De Angeli (2014a); Miniukovich and De Angeli (2014b); Miniukovich and De Angeli (2015a); Miniukovich and De Angeli (2015b); Miniukovich et al. (2018); Reinecke et al. (2013); Reinecke and Gajos (2014); Riegler and Holzmann (2018); Uribe et al. (2017)
	Complexity	Miniukovich and De Angeli (2014b); Miniukovich and De Angeli (2015a); Miniukovich et al. (2018); Ngo et al. (2000); Ngo and Byrne (2001); Purchase et al. (2011); Reinecke et al. (2013); Reinecke and Gajos (2014); Riegler and Holzmann (2018); Zen and Vanderdonckt (2014)
	Concentricity	Zen and Vanderdonckt (2014)
	Color contrast	Maity et al. (2015); Maity and Bhattacharya (2019); Miniukovich and De Angeli (2014a); Miniukovich and De Angeli (2015a)
	Density	Bessghaier et al. (2021); Maity et al. (2015); Mbenza and Burny (2020); Ngo and Byrne (2001); Purchase et al. (2011); Riegler and Holzmann (2018); Zen and Vanderdonckt (2014)
	Economy	Mbenza and Burny (2020); Ngo and Byrne (2001); Purchase et al. (2011); Zen and Vanderdonckt (2014)
	Edge congestion	Miniukovich and De Angeli (2014a); Miniukovich and De Angeli (2014b); Miniukovich and De Angeli (2015a)
	Element ratio	Maity and Bhattacharya (2019)

Equilibrium	Maity et al. (2015); Ngo et al. (2000); Ngo and Byrne (2001); Purchase et al. (2011); Reinecke et al. (2013); Zen and Vanderdonckt (2014)
Grid	Miniukovich and De Angeli (2014b); Miniukovich and De Angeli (2015a)
Grouping	Mbenza and Burny (2020) ; Miniukovich and De Angeli (2014b)
Homogeneity	Bessghaier et al. (2021); Mbenza and Burny (2020); Ngo and Byrne (2001); Purchase et al. (2011); Zen and Vanderdonckt (2014)
Hue	Maity et al. (2015); Maity and Bhattacharya (2019); Meier (1988)
Lighting	Maity et al. (2015); Uribe et al. (2017)
Number of elements	Maity et al. (2015); Reinecke et al. (2013); Reinecke and Gajos (2014)
Order	Ngo et al. (2000); Ngo and Byrne (2001); Purchase et al. (2011); Zen and Vanderdonckt (2014);
Proportion	Mbenza and Burny (2020); Ngo and Byrne (2001); Purchase et al. (2011); Zen and Vanderdonckt (2014); Zen and Vanderdonckt (2016)
Regularity	Bessghaier et al. (2021); Ngo and Byrne (2001); Zen and Vanderdonckt (2014)
Rhythm	Maity et al. (2015); Ngo and Byrne (2001); Zen and Vanderdonckt (2014)
Saturation	Maity et al. (2015); Maity and Bhattacharya (2019); Meier (1988)
Sequence	Bessghaier et al. (2021); Ngo and Byrne (2001); Purchase et al. (2011); Zen and Vanderdonckt (2014)
Sharpness	Maity and Bhattacharya (2019)
Simplicity	Bessghaier et al. (2021); Mbenza and Burny (2020); Ngo and Byrne (2001); Purchase et al. (2011); Zen and Vanderdonckt (2014); Zen and Vanderdonckt (2016)
Smoothness	Maity et al. (2015); Maity and Bhattacharya (2019)
Symmetry	Maity et al. (2015); Mbenza and Burny (2020); Miniukovich and De Angeli (2014a); Miniukovich and De Angeli (2014b); Miniukovich and De Angeli (2015a); Miniukovich et al. (2018); Ngo et al. (2000); Ngo and Byrne (2001); Reinecke et al. (2013); Reinecke and Gajos (2014); Zen and Vanderdonckt (2014); Zen and Vanderdonckt (2016)
Texture	Miniukovich and De Angeli (2015b); Uribe et al. (2017)
Unity	Bessghaier et al. (2021); Mbenza and Burny (2020); Ngo and Byrne (2001); Purchase et al. (2011); Zen and Vanderdonckt (2014)
Value	Maity and Bhattacharya (2019)
White space	Maity and Bhattacharya (2019); Miniukovich and De Angeli (2015a)
Metrics related to typography	
Contrast (luminance, chromatic)	Maity et al. (2016); Maity and Bhattacharya (2019)
Font size	Maity et al. (2016); Maity and Bhattacharya (2019)
Letter spacing	Maity et al. (2016); Maity and Bhattacharya (2019)
Line height	Maity et al. (2016); Maity and Bhattacharya (2019)
Word spacing	Maity et al. (2016); Maity and Bhattacharya (2019)

Table 2 illustrates the frequencies of different metric-based aspects. The most studied metrics include symmetry, color and complexity, as well as balance. The least

frequent metrics are brightness and value. Automation of GUI aesthetics has been argued to be precise, cost-efficient and fast (Ngo et al. 2000; Zen and Vanderdonckt 2014). However, the metric aspects remain scattered and thus, different configurations and controversial results of metric evaluations occur which hinders systematic automation of GUI aesthetics evaluation and design.

Table 3. Dimensions of UI aesthetics based on perceptions

Type	Dimension	Studies
Evaluation based on human perceptions	Hedonic quality–identification, Hedonic quality–stimulation, Pragmatic quality	Hassenzahl et al. (2003); Hassenzahl (2004)
	Excellence/Inferiority, Graciousness/Harshness, Idleness/Liveliness, Normalness/Bizarreness, Complexity/Simplicity	Jylhä and Hamari (2020)
	Classical and Expressive Aesthetic value	Lavie and Tractinsky (2004)
	Simplicity, Diversity, Color, Craftsmanship	Moshagen and Thielsch (2010); Moshagen and Thielsch (2013)
	Aesthetic dimensions of websites	Park et al. (2004);
	Use of color, Symmetry/aesthetic style, Structured layout, Depth of field, Choice of media, People and Personality, Unusual images	Sutcliffe (2002)

Table 3 summarizes aesthetic dimensions examined by the evaluation methods based on human perceptions. Similar to the metric aspects, the qualitative instruments also introduce several different concepts and scales, i.e., semantic differential (Hassenzahl et al. 2003; Hassenzahl 2004; Jylhä and Hamari 2020) and Likert scales (Lavie and Tractinsky 2004; Moshagen and Thielsch 2010; 2013; Park et al. 2004; Sutcliffe 2002). However, overlapping themes include studying color, simplicity and complexity, quality of aesthetic design, as well as uniqueness. The benefits of these methods are the consideration of the subjective experience (Hassenzahl et al. 2003; Jylhä and Hamari 2020), nevertheless, administering a survey can be time-consuming with a risk of a small and/or non-diverse pool of participants, which may result in poor generalizability of research.

Table 4. Aspects based on deep learning and eye-tracking

Type	Aspects	Studies
Deep learning	Clustering	Wu et al. (2016)
	Feature extraction	Dou et al. (2019); Khani et al. (2016); Liu and Yiang (2021); Wu et al. (2011); Wu et al. (2016); Xing et al. (2021)
	Feature dimension reduction	Khani et al. (2016)
	HSV color assessment	Dou et al. (2019); Liu and Yiang (2021); Wu et al. (2011); Wu et al. (2016)
	Regression mapping	Dou et al. (2019); Wu et al. (2016); Xing et al. (2021)
	SVM classification	Khani et al. (2016); Liu and Yiang (2021); Wu et al. (2016)
	Text scoring	Liu and Yiang (2021); Wu et al. (2011); Wu et al. (2016)
	Texture	Wu et al. (2016); Xing et al. (2021)
	Transfer learning	Dou et al. (2019)
	Eye-tracking	Areas of gaze interest (AOI)
Gaze duration		Gu et al. (2020); Pappas et al. (2020)
Gaze fixation		Gu et al. (2020); Pappas et al. (2020)
Pupil diameter		Pappas et al. (2020)
Saccade		Gu et al. (2020); Pappas et al. (2020)

Table 4 presents aspects related to deep learning and eye-tracking that predict GUI aesthetics. These measurement types are relatively novel and thus scarce, however, promising results have been provided by the extant corpus on the benefits of these methods. Convolutional neural networks have been argued to be effective for GUI aesthetic evaluation because the data sample is usually extensive, and quick, reliable results have been acquired when compared to human ratings (Dou et al. 2019; Khani et al. 2016; Xing et al. 2021). Eye-tracking on the other hand is said to offer new ways for design that will take into account gaze behavior in an unobtrusive manner and will be able to inform researchers and designers about perceptions of visual aesthetics (Pappas et al. 2020). However, as a new field of research methodology, further studies are required to confirm the validity of these approaches.

2.4 Use contexts of the measurement instruments

Table 5 describes the GUI elements presented in prior literature. They were categorized as they appear, namely as the entire GUI, separate objects (i.e., windows, menus and icons) as well as images (i.e., geometric arrangements of different sizes and shapes, layout skins), typography or video. Some studies include various element types. This has been considered and marked accordingly as seen in Table 5.

Table 5. Elements of graphical user interface aesthetics

Type	Studies
Entire GUI	Bessghaier et al. (2021); Dou et al. (2019); Gu et al. (2020); Khani et al. (2016); Lavie and Tractinsky (2004); Mbenza and Burny (2020); Miniukovich and De Angeli (2014a); Miniukovich and De Angeli (2014b); Miniukovich and De Angeli (2015a); Miniukovich and De Angeli (2015b); Miniukovich et al. (2018); Moshagen and Thielsch (2010); Moshagen and Thielsch (2013); Ngo et al. (2000); Ngo and Byrne (2001); Pappas et al. (2020); Park et al. (2004); Reinecke et al. (2013); Reinecke and Gajos (2014); Riegler and Holzmann (2018); Sutcliffe (2002); Uribe et al. (2017); Wu et al. (2011); Xing et al. (2021); Zen and Vanderdonckt (2014); Zen and Vanderdonckt (2016)
Windows, icons and/or menus	Meier (1988); Jylhä and Hamari (2020); Wu et al. (2016)
Images	Hassenzahl et al. (2003); Hassenzahl (2004); Liu and Yiang (2021); Maity et al. (2015); Maity and Bhattacharya (2019); Purchase et al. (2011); Wu et al. (2016)
Text and typography	Liu and Yiang (2021); Maity et al. (2016); Maity and Bhattacharya (2019); Purchase et al. (2011); Wu et al. (2016)
Video	Liu and Yiang (2021)

As shown here, the majority of prior research evaluated graphical user interfaces as an entity. While layouts are to be designed in such a way that different elements work seamlessly together, prior literature (Maity et al. 2016; Zen and Vanderdonckt, 2016; Xing et al. 2021) suggests that contradictory results in evaluations may be caused by analyzing whole user interfaces without considering the content. For example, designing buttons is different from defining typefaces (Maity et al. 2016). Layout designs vary, which may cause difficulties in generalization (Jylhä and Hamari, 2020). This point of GUI aesthetics evaluation is still open for discussion and requires further research.

Table 6 presents the use context in which the measurement instruments have been applied. The contexts include *desktop* (e.g., website, software) and *mobile*

interfaces. In some studies, use context differed from the two main categories, thus they were classified as *other*. Some studies also included multiple contexts, which has been considered accordingly in Table 6.

Table 6. Use contexts for measuring graphical user interface aesthetics

Context type	Interface	Studies
Desktop	Website	Dou et al. (2019); Gu et al. (2020); Khani et al. (2016); Lavie and Tractinsky (2004); Liu and Yiang (2021); Maity et al. (2016); Maity and Bhattacharya (2019); Miniukovich and De Angeli (2014a); Miniukovich and De Angeli (2015a); Miniukovich and De Angeli (2015b); Miniukovich et al. (2018); Moshagen and Thielsch (2010); Moshagen and Thielsch (2013); Pappas et al. (2020); Park et al. (2004); Purchase et al. (2011); Reinecke et al. (2013); Reinecke and Gajos (2014); Sutcliffe (2002); Uribe et al. (2017); Wu et al. (2011); Wu et al. (2016); Xing et al. (2021); Zen and Vanderdonckt (2014); Zen and Vanderdonckt (2016)
	Software or application	Mbenza and Burny (2020); Meier (1988); Ngo et al. (2000); Ngo and Byrne (2001)
Mobile		Bessghaier et al. (2021); Jylhä and Hamari (2020); Miniukovich and De Angeli (2014b); Miniukovich and De Angeli (2015a); Miniukovich and De Angeli (2015b); Riegler and Holzmann (2018)
Other (i.e., MP3 layout skins, images)		Hassenzahl et al. (2003); Hassenzahl (2004); Miniukovich et al. (2018)

Website evaluation is distinctly the most common use context for GUI aesthetics assessment in instrument development. Nevertheless, the context types in which the measures have been applied are not exclusive, as several studies encourage use the suggested measurement models in other GUI contexts as well (e.g., Hassenzahl et al. 2003; Hassenzahl 2004; Jylhä and Hamari 2020; Mbenza and Burny 2020; Zen and Vanderdonckt 2014; Zen and Vanderdonckt 2016; Xing et al. 2021). Evaluation methods for mobile GUI aesthetics are the second most common despite the vast difference in numbers compared to website evaluation methods. This is likely to change in the future as mobile devices are getting increasingly popular. Users have been reported to perceive visual aesthetics similarly in these two contexts (Miniukovich & De Angeli 2014b; 2015a), however, there is a lack of research concerning further comparison.

2.5 Validity of the measurement instruments

Measurement instruments have been developed, applied and tested by various means, but discussion about the suitability of the instruments considering their reliability and validity is missing. Therefore, it is of importance to investigate and categorize the validity of measurement instruments that have been introduced by prior studies.

In prior research, some evaluation methods have not been validated, i.e., the degree to which an instrument measures what it claims to measure (Leedy and Ormrod 2004) has not been confirmed. Therefore, non-validated measurement instruments were not included in this section as a method is ideally robust and reliable. This means that only those studies whose measurement properties (i.e., reliability and validity) have been assessed using standardized criteria were chosen for this section. Table 7 lists studies according to their validity status, divided first into the primary measurement types and further categorized as *validated* and *non-validated* instruments.

Table 7. Validated and non-validated instruments

Type	Status	Studies
Evaluation based on graphical features	Validated	Bessghaier et al. (2021); Maity et al. (2015); Maity et al. (2016); Maity and Bhattacharya (2019); Miniukovich and De Angeli (2014a); Miniukovich and De Angeli (2014b); Miniukovich and De Angeli (2015a); Miniukovich et al. (2018); Ngo and Byrne (2001); Reinecke et al. (2013); Reinecke and Gajos (2014); Uribe et al. (2017); Zen and Vanderdonckt (2014); Zen and Vanderdonckt (2016)
	Non-validated	Mbenza and Burny (2020); Meier (1988); Miniukovich and De Angeli (2015b); Ngo et al. (2000); Riegler and Holzmann (2018); Purchase et al. (2011)
Evaluation based on human perceptions	Validated	Jylhä and Hamari (2020); Lavie and Tractinsky (2004); Moshagen and Thielsch (2010); Moshagen and Thielsch (2013); Park et al. (2004)
	Non-validated	Hassenzahl et al. (2003); Hassenzahl (2004); Sutcliffe (2002)
Other (i.e., deep learning, eye-tracking)	Validated	Dou et al. (2019); Khani et al. (2016); Pappas et al. (2020); Wu et al. (2011); Wu et al. (2016); Xing et al. (2021)
	Non-validated	Gu et al. (2020); Liu and Yiang (2021)

Overall, most instruments were validated as opposed to non-validated studies. The standard means to investigate the accuracy of instruments based on graphical features is usually by correlation or regression analyses to fit subjective ratings of aesthetics using the measures. Other means include t-tests, interrater reliability and interrater agreement. Many of the metric-based approaches are based on the model proposed by Ngo et al. (2000). The model was later validated by Ngo and Byrne (2001) via multiple regression analyses, reporting high correlations between computed aesthetic value and the aesthetics ratings of design experts. These results were replicated only to an extent during the development of QUESTIM (Zen and Vanderdonckt 2014; 2016), who obtained medium degree of interrater agreement and low reliability for calculating symmetry and balance, after which a new formula for balance is introduced. The validation of ADDET (Bessghaier et al. 2021) was constructed via t-tests on the basis of QUESTIM (Zen and Vanderdonckt 2014; 2016) and was found to be more effective in estimating aesthetic value. However, ADDET is aimed at mobile user interfaces while QUESTIM is directed at all platforms. Other metric measures aside from the model by Ngo et al. (2000) and Ngo and Byrne (2001) that were validated via regression analyses (Miniukovich and De Angeli 2014a; 2014b; 2015a; Reinecke et al. 2013; Reinecke and Gajos 2014; Uribe et al. 2017) indicated that the variability of visual aesthetics explained approximately 50% of aesthetic preference in GUI evaluation. Additionally, interrater reliability (Reinecke et al. 2013; Reinecke and Gajos 2014; Miniukovich and De Angeli 2014a; 2014b; 2015a; Miniukovich et al. 2018) and interrater agreement (Zen and Vanderdonckt 2016) analyses confirmed high reliability among scores. However, Reinecke and Gajos (2014) noted that aesthetic preferences significantly differ according to demographic factors. Studies that reported the metrics validity via RMSE (Maity et al. 2015; 2016; Maity and Bhattacharya 2019) found that text aesthetics could be measured by the score of 0.58 and interface aesthetics by 0.79. Although these quantified measures seem to produce relatively high accuracy scores in validity results, the configurations of metrics are dispersed, and thus, their performance is difficult to predict in different use contexts. Theoretically, it can be postulated that because the metrics by Ngo et al. (2000) and Ngo and Byrne (2001) have been previously examined by various studies with satisfactory results, it benefits the measurement instrument in terms of suitability. Nevertheless, the results have not been optimal which indicates that the measure requires further development. This applies also for the other metric instruments that have not been as widely used.

The validity of instruments based on human perceptions is commonly investigated by confirmatory factor analysis (CFA) and model fit indices such as Chi

square test (χ^2), comparative fit index (CFI), root mean square error of approximation (RMSEA), and standardized root mean square residual score (SRMR). Additionally, average variance extracted (AVE) and composite reliability (CR) have been examined. Comparing the validity of the survey-based instruments presented in Table 8, VISQUAL (Jylhä and Hamari 2020) reported support for discriminant validity, however, convergent validity and composite reliability remained open for critique. The Classical and Expressive Aesthetics Scale (Lavie and Tractinsky 2004) indicated that reliability tests, factor structure and validity reflect the aesthetics scales adequately. VisAWI (Moshagen and Thielsch 2010; Moshagen and Thielsch 2013) successfully demonstrated convergence validity but lacked in divergent validity. While the scale by Park et al. (2004) was found to have high convergent and discriminate validity of the 13 aesthetic dimensions and 30 aesthetic adjectives, the study seems somewhat lacking in instructions on how to use the measurement instrument. Moreover, the scale includes the most items, which may result in a heavy task for participants and a time-consuming process in general. Validation is a cumulative, on-going process with multiple methods and samples (Spector 1992). Both the Classical and Expressive Aesthetics Scale (Lavie and Tractinsky 2004) and VisAWI (Moshagen and Thielsch 2010; Moshagen and Thielsch 2013) have been used in future studies, enforcing the reliability and validity of the measures. As such, being a novel method, VISQUAL (Jylhä and Hamari 2020) is still to be tested further in other contexts.

Other instruments based on deep learning and eye-tracking are validated by observing prediction accuracy and error rates by e.g., cross-validation, the residual sum of squares error (RSSE) and the root-mean-square error (RMSE), to determine how much the results differ from human perceptions. Two studies (Dou et al. 2019; Khani et al. 2016) using the same dataset with different CNN approaches both tested square errors, where Dou et al. (2019) reached a lower test error rate, obtaining better results. The eye-tracking study by Pappas et al. (2020) used normalized RMSE (NRMSE) to measure accuracy prediction, concluding the analysis with high accuracy results. RSSE was used by Wu et al. (2011; 2016) to measure regression performance, as well as to compare error rates between four other studies, where the scores of Wu et al. (2016) were the lowest. Xing et al. (2021) validated their model by examining error rates (MSE) of several DCNNs to find that SE-VGG19 was the most effective in predicting aesthetic preference. As machine learning is still a relatively new phenomenon, the measures are still in their exploration state and comparison may be difficult due to different approaches and means of validity analyses. Regardless, Dou et al. (2019) denote that their results outperform previous

similar methods indicating efficiency for aesthetics evaluation, which shows promise for the development of deep learning and aesthetic evaluation in general.

2.6 Contradictions and gaps in prior research

Perhaps due to the relative novelty of this research topic, contradictions in general consensus exists, that leads into identifiable research gaps. This dissertation intends to rationalize some of the inconsistencies in prior literature and fill the gaps found during the course of the whole of this work. Prior literature (Maity et al., 2015; Maity et al., 2016; Zen and Vanderdonckt, 2016) suggests that contradictions in metric-based evaluation theories aesthetics in GUI research are perhaps caused by analyzing user interfaces as entities. This gap 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). Moreover, empirical studies on GUI aesthetics have often relied on website layouts or skins as study objects (Hassenzahl, 2004). This can be problematic, as measuring perceived attractiveness of layouts or skins does not necessarily reveal which elements in the user interface are successful. This is a shortcoming of the empirical measurements as inclusion may prevent calculating genuine values of user interfaces. Prior study (Vanderdonckt and Gillo, 1994) that automated 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. 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, use context as well as measurement validity, hence, further comparison between measurement instruments is needed in order to provide effective tools for both academic and industry use. The publications in this dissertation fill these gaps firstly by mapping out, systematically categorizing, and comparing measurement instruments by prior literature in order to clarify use contexts of the tools available (publication 1). Secondly, the contradiction in evaluating GUIs as entities instead of single elements or by using layouts and skins as study material is countered by the development and validation of the measurement instrument VISQUAL used to assess the aesthetic features of graphical assets as separate or as an entity, by utilizing a varied set of GUI icons as study material (publication 2). Thirdly, user perceptions,

a major part of attractive graphical user interface design, are observed as an attempt to create design guidelines for GUI elements (publication 3) based on VISQUAL, further validating the tool. Finally, use context and demographic factors are studied (publication 4) to advance the design of graphical user interface aesthetics according to target group.

3 METHODS AND DATA

The objectives of the publications in this dissertation were to map out how the visual aesthetics of graphical user interfaces can be modelled, evaluated, and assessed; what is the dimensional structure of aesthetics perceptions of graphical user interface elements; as well as how perceptions of different users predict the attractiveness of graphical user interface elements. The methodology therefore builds on an extensive research approach that is first described before diving into detail about the methods in each publication. An overview of the methodology includes a systematic literature review for publication 1, and a large-scale online survey experiment with a psychometric instrument (VISQUAL) for publications 2, 3 and 4 that were utilized to further investigate the aforementioned research aims. These methods were deemed suitable for the studies to ensure validity and generalizability of the results. Therefore, the following sections describe the overall approach, review and experiment processes, as well as participants, materials, measurements, procedure and analysis methods. Parts of this chapter have been previously published in peer-reviewed journal articles (Jylhä and Hamari 2019, 2020, 2021).

3.1 Research approach

The research approach in the whole of this dissertation consists of steps starting from broad impressions narrowed down to detailed methods of data collection, analyses, and interpretation of the findings. The paradigm was to examine graphical user interface aesthetics, their formulation and user impressions, requiring both theoretical assessment as well as participant data. Due to the complex nature of the research questions, the approach of data collection was countered with two types of standardized quantitative methods: a systematic literature review and a large-scale online survey experiment. This was to ensure data from multiple sources. Employing quantitative methodology requires an understanding of deductive research and statistical data analysis that investigates relationships with variables as well as descriptive values (Trochim, 2000). Several measures for data analysis were used in the publications, and they are outlined in the sections following the detailed

introductions of each approach. Quantitative research was chosen as the approach for the publications in this dissertation as it can be deemed objective in nature compared with qualitative methods. However, choosing one approach over the other has been limits the scope of the study (Creswell and Clark, 2011). This has been noted as one of the limitations of this dissertation. Future research is invited to study the topics presented in the publications of this dissertation with qualitative methods in order to add to the findings from a more subjective perspective.

The research was conducted in accordance with all relevant ethical guidelines of the Finnish Advisory Board of Research Integrity TENK (Kohonen et al., 2019). All respondents participated in the study voluntarily. Participants received information on the study objective, practical steps, processing of personal data, and their rights to discontinue participation or withdraw their consent to participate at any time with no negative consequences. Personal data gathered during the research was processed following the requirements of General Data Protection Regulation GDPR (Voigt and Von dem Bussche, 2017) and analyzed without personal information. Access to research data was limited to the researchers and the personal data was removed from the research data as soon as it was not needed for research purposes.

3.2 Literature review

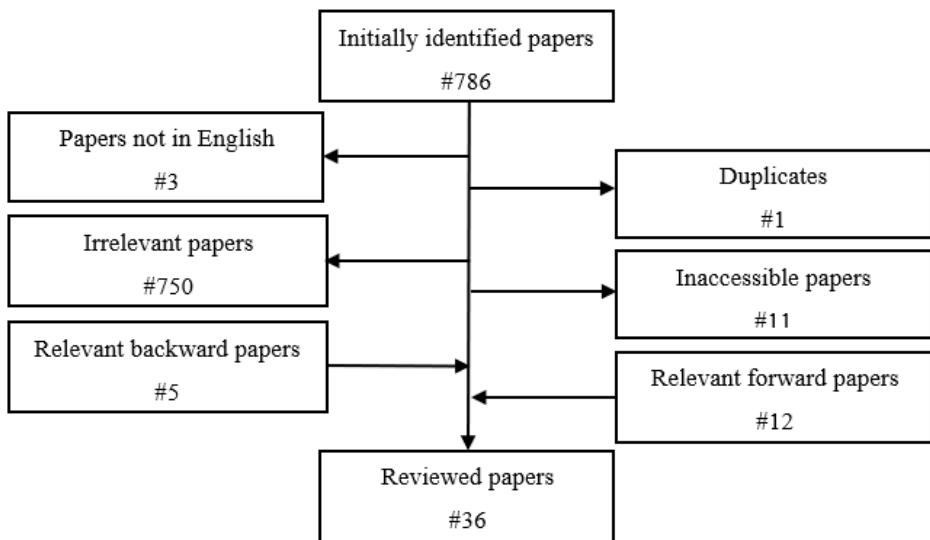
The literature review in publication 1 followed a summarization of knowledge protocol (Paré et al., 2015) that includes narrative, descriptive and scoping reviews that aim to map and describe a body of literature. In order to organize the literature by publications and concepts, a concept-centric coding strategy (Webster and Watson, 2002) was adopted to describe the corpus quantitatively. As such, the review process was divided in the following stages: 1) explorative literature search to map relevant keywords, 2) systematic literature search, 3) inclusion and exclusion procedures, 4) backward search, 5) forward search, 6) concept-centric coding and analysis of literature, and 7) reporting of findings.

Based on the keyword mapping, the search string contained three main keywords with related terms: *graphical user interface* (e.g., GUI, user interface), *aesthetics* (e.g., aesthetic, appearance, beauty), and *method* (e.g., measurement, metric, survey, evaluation, assessment). Thus, the following search string was used in this study:

((TITLE-ABS-KEY (gui OR "graphical interface" OR "user interface" OR ui) AND TITLE-ABS-KEY (aesthetic* OR appearance* OR beauty OR beautiful OR attractive*) AND TITLE-ABS-KEY (measure* OR method* OR assess* OR evaluat* OR metric* OR questionn* OR survey OR scale* OR framework* OR instrument* OR heuristic* OR guid*)) AND (EXCLUDE (SUBJAREA , "ENGI") OR EXCLUDE (SUBJAREA , "PHYS") OR EXCLUDE (SUBJAREA , "MATH") OR EXCLUDE (SUBJAREA , "MEDI") OR EXCLUDE (SUBJAREA , "EART") OR EXCLUDE (SUBJAREA , "BIOC") OR EXCLUDE (SUBJAREA , "AGRI") OR EXCLUDE (SUBJAREA , "ENER") OR EXCLUDE (SUBJAREA , "CHEM") OR EXCLUDE (SUBJAREA , "CENG") OR EXCLUDE (SUBJAREA , "ENVI") OR EXCLUDE (SUBJAREA , "MATE") OR EXCLUDE (SUBJAREA , "PHAR") OR EXCLUDE (SUBJAREA , "HEAL") OR EXCLUDE (SUBJAREA , "NURS") OR EXCLUDE (SUBJAREA , "DENT") OR EXCLUDE (SUBJAREA , "VETE") OR EXCLUDE (SUBJAREA , "IMMU"))) AND (EXCLUDE (DOCTYPE , "cr") OR EXCLUDE (DOCTYPE , "no") OR EXCLUDE (DOCTYPE , "re") OR EXCLUDE (DOCTYPE , "bk"))).

Scopus was selected as the search engine due to its large coverage of peer-reviewed literature in relevant fields. The initial search was performed in August 2021, and it yielded 786 results in total. The literature search process is shown in Figure 2.

Figure 2. Literature search process and outcomes



The studies were screened according to the following selection criteria to make the results more precise: 1) studies written in English, 2) non-duplicate studies (i.e., with the same DOI), 3) studies available for download, and 4) studies that propose, analyze, and/or apply measurement methods for the visual aesthetics of graphical user interfaces. Once applied in the retrieved studies, only 19 of them met the selection criteria. Following the backward snowballing process to identify additional work, 5 relevant papers were found. Forward references revealed 12 more. This resulted in a total of 36 manuscripts.

An obviously large number of studies was excluded due to being labelled as irrelevant. As the terms in the search string are commonly in use in other fields such as physics, medical science and engineering, the aim was to limit the search by subject area in the query while maximizing the inclusion of prior scattered measures. Thus, the query was challenging to construct as more exclusive. The coding process therefore entailed careful manual labor to rule out publications that were out of the scope of publication 1. Additionally, some studies were encountered that were in the field of HCI and presented theories or tools for user interfaces, being relevant in that sense. However, further investigation revealed that the instruments were targeted towards usability with aesthetics being only a minimal or nonexistent part of the study (e.g., Alemerien and Magel, 2015; Yang and Klemmer, 2009).

3.3 Online survey and vignette experiment

An online survey was selected as the data collection method for publications 2, 3 and 4, as it enabled a large-scale examination of user perceptions on visual qualities of graphical user interface elements, as well as measurement validity and reliability. The study setting was a within-subjects vignette experiment. Participants were assigned to evaluate 4 randomized icons from a total of pre-selected 68 game app icons across 4 categories (concrete, abstract, character and text) in a hypothetical situation setting instead of a description more typical to vignette studies. 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. Two participants were raffled to receive a prize (Polar Loop 2 Activity Tracker). No other participation fees were paid. Participants were informed the purpose of the study and assured that they will remain anonymous during the experiment and data analysis.

3.3.1 Participants

A nonprobability convenience sample was composed initially of 569 respondents who each assessed 4 game app icons through a survey-based randomized within-subjects vignette experiment, resulting in 2276 icon evaluations. A within-subjects approach was chosen as opposed to between-subjects approach in order to expose each participant to all conditions (i.e., 4 icon evaluations by category) of the experiment. This dataset was used in full for publications 2 and 3. For publication 4 on demographics, 15 responses without identifiable gender were removed due to insufficient representation. Additionally, 41 responses from older age groups were identified as outliers and removed, resulting in a total of 513 respondents. Please refer to Table 8 for full demographic details of participants.

Table 8. Demographic information

		n	%
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
	Higher than master's degree	9	1.6
Employment	Working full-time	133	23.4
	Working part-time	62	10.9
	Student	351	61.7
	Unemployed	11	1.9
	Retired	1	.2
Gender	Male	297	52.2
	Female	257	45.2
	Other	15	2.6
Yearly income	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



The majority of the participants was from Finland (92.8%). Only slightly more than half of the sample body was 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%).

3.3.2 Materials

Sixty-eight game app icons from Google Play Store were selected for the study. The decision to narrow down the sample to game app icons was made to eliminate further variability that might stem from the nature of the app and thus increase internal validity of the experiment, but also external validity in terms of results applied to the game icons. In order to avoid any systematic bias, 4 icons corresponding to dominant icon styles (concrete, abstract, character and text) were selected from each of 17 categories for game apps (action, adventure, arcade, board, card, casino, casual, educational, music, puzzle, racing, role playing, simulation, sports, strategy, trivia and word). Because icon design for app stores is category-dependent (Shu and Lin, 2014), we considered it justified to include icons from all categories. Prior literature highlights the relevance of concreteness and abstractness in icon design (e.g., 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; McDougall et al., 2000; Moyes and Jordan, 1993; Rogers and Osborne, 1987), hence they were included in this experiment. Looking at the icons on app stores, characters and typography are prevalent elements usually seen on app icons. It has been argued that faces on app icons are widely used because of the immediate impact and memorability they have due to neural processing of facial expressions (Chartboost, 2015). Furthermore, as the study design is based on prior research (Shaikh, 2009) on onscreen typeface and usage, text elements were included. During the selection phase we ensured that one icon from each category was dominantly characteristic of one of these 4 attributes. Please refer to Table 9 for the icons.

Table 9. Icons used in the experiment

Category	Concrete	Abstract	Character	Text
Action				
Adventure				
Arcade				
Board				
Card				
Casino				
Casual				
Educational				
Music				
Puzzle				
Racing				
Role Playing				
Simulation				

Sports				
Strategy				
Trivia				
Word				

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 3rd to 17th 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 as visually rich a sample of icons as possible, meaning that several different computer graphic techniques were included, such as 2D and 3D rendered images.

3.3.3 Measurements

Semantic differential scale was used to measure respondent evaluations of aesthetic aspects of the icons. A total of 22 adjective pairs was formulated and assigned to each icon. The polarity of the adjective pairs was reversed 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.

All of the adjective pairs were chosen according to prior research (Shaikh, 2009) on onscreen typeface design and usage. Additionally, adjectives related to icons were added as suggested per previous literature on effective icon design. 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; McDougall et al., 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 were added to specifically measure the aesthetic properties of the icons include *professional and unprofessional*, *colorful and colorless*, *realistic and unrealistic* as well as *two-dimensional and three-dimensional* (Jylhä and Hamari, 2019).

Developed further into a five-factor model entitled VISQUAL (Jylhä and Hamari, 2020), an instrument for measuring visual qualities of graphical user interface elements, the scale was used to observe underlying latent constructs in this study. VISQUAL consists of the aforementioned adjective pairs that were further divided into the following dimensions: *Excellence/Inferiority*, *Graciousness/Harshness*, *Idleness/Liveliness*, *Normalness/Bizarreness* and *Complexity/Simplicity*. Table 10 lists the VISQUAL constructs and adjective pairs.

Table 10. Constructs, means and standard deviations (adjusted items bolded)

Factor	Adjective pair	Mean	Std.
Excellence/ Inferiority	Good–Bad	4.34	1.641
	Professional–Unprofessional	4.22	1.736
	Beautiful–Ugly	4.57	1.618
	Expensive–Cheap	4.83	1.563
	Strong–Weak	3.93	1.464
Graciousness/ Harshness	Soft–Hard	3.81	1.545
	Relaxed–Stiff	4.47	1.560
	Masculine–Feminine	4.34	1.388
	Delicate–Rugged	4.42	1.368
	Happy–Sad	3.80	1.507
	Colorful–Colorless	3.77	1.810
Idleness/ Liveliness	Warm–Cool	3.97	1.436
	Fast–Slow	3.87	1.576
	Quiet–Loud	4.12	1.601
	Exciting–Calm	3.96	1.452
	Active–Passive	3.97	1.708
Normalness/ Bizarreness	Young–Old	3.98	1.611
	Concrete–Abstract	4.03	1.998
	Realistic–Unrealistic	4.22	1.592
Complexity/ Simplicity	Ordinary–Unique	4.60	1.651
	Simple–Complex	4.69	1.669
	Three-dimensional–Two-dimensional	3.33	1.863

Two versions of the model exist, the initial model with 22 adjective pairs and an adjusted model of 15 adjective pairs. In Table 10, the bolded adjective pairs represent those included in the adjusted model of 15 adjective pairs. Table 10 also presents an overview of the means and standard deviations. There were no outlier values and the range between the lowest and highest scores clustered closely to the average even though the 68 icons were quite different from each other. All the mean scores were between 3.5 and 4.5 for each evaluation. This indicates little skewness in the data.

Additional to the semantic scales, a seven-point Likert scale was utilized to measure the degree of disagree-agreement of the respondents with respect to the likelihood of them clicking, downloading, and purchasing the imagined app behind the icon with an instruction title: “Overall evaluation (judging by the icon alone)” followed by questions: “Compared to the mobile game icons I usually click, I would click this icon”, “Compared to the icons of mobile games I usually download, I would click this icon” and “Compared to the icons of mobile games I usually purchase, I would click this icon.” Respondents were provided the following options on the seven-point scale: “Strongly disagree”, “Disagree”, “Somewhat disagree”, “Neither agree nor disagree”, “Somewhat agree”, “Agree” and “Strongly agree”. Moreover, respondents were asked to give an overall evaluation score for the design of each icon by grading them on a seven-point scale to further assess consumer perceptions of icon successfulness.

3.3.4 Procedure

The data was collected through a randomized survey-based online 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. Last, demographic information (age, gender, etc.) was asked.

The survey took about 10 minutes to complete. The survey was implemented via Surveygizmo, an online survey tool. All content was in English. The data was analyzed with IBM SPSS Statistics and Amos version 24 as well as Microsoft Office Excel 2016.

3.4 Analyses

3.4.1 Structural equation modelling

As the aim of publication 2 was to assess the latent psychometric properties of the measurement instrument VISQUAL, which was also used to collect data for publications 3 and 4 respectively, covariance-based structural equation modelling (CB-SEM) was utilized. SEM was deemed suitable for this purpose, as it is intended for investigating models consisting of complex relationships between multiple latent variables. SEM is a valid method especially in cases where the objective is to perform theory testing combining factor analyses and multiple linear regression (Hair et al., 2016) followed by validity calculations to produce a consolidated scale.

3.4.2 Exploratory and confirmatory factor analysis

As a part of SEM, exploratory factor analysis (EFA) and confirmatory factor analyses (CFA) were performed to develop and validate the measurement instrument VISQUAL. First, EFA with Varimax rotation and Kaiser normalization was performed to explore factor structures of the 22 adjective pairs listed previously in Table 4. There were no initial expectations regarding the number of factors. Principal component analysis (PCA) was used as extraction method to maximize the variance extracted. The analysis revealed 5 dimensions introduced previously in Table 4.

CFA was then performed in two consecutive stages including the initial model with 22 adjective pairs as well as with an adjusted model of 15 adjective pairs to test whether the proposed theory could be applied to similar latent constructs. CFA consists of model fit tests as well as validity and reliability analysis. For this purpose, as per recommendation by 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 than $.10$ (Kenny, 2020).

Following the model fit tests, the scale was examined for validity and reliability by standardized measures, namely Cronbach's alpha, average variance extracted (AVE) and composite reliability (CR). Prior literature suggests 0.7 as the typical cut-off level for acceptable values for Cronbach's alpha (Nunnally and Bernstein 1994). Convergent validity is determined by AVE that is greater than $.5$, and composite reliability is attained by values greater than $.7$ (Hair et al., 2016). In terms of discriminant validity, the maximum shared variance (MSV) should be less than the AVE, and the square root of the AVE of each construct should be larger than any correlation between the same construct and all the other constructs (Fornell and Larcker 1981).

The initial model fit tests as well as validity and reliability analyses showed that the instrument was not an optimally fitting measurement model, with additional problems relating to unacceptable item loadings in the CFA (standardized weights). Loadings should fall between $.32$ and 1.00 (Matsunaga 2010; Tabachnick and Fidell

2007), which indicated the need for post-hoc adjustments. After the removal of poorly behaved reflective indicators as recommended by prior literature (Brown 2015; MacKenzie et al. 2011), the overall model fit improved. Furthermore, examining strong modification indices (MI = 3.84) and covarying items accordingly (MacKenzie et al. 2011) proved 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 were significantly improved.

3.4.3 Regression analysis

Multiple linear regression analysis (MLRA) was used in publication 3 and 4 to predict the outcome of the variables. Prior to the analyses, multicollinearity tests were performed with variance inflation factors (VIF). No critical levels of multicollinearity were found between the variables. In publication 3, MLRAs were performed to examine the relationships between user perceptions of GUI elements (i.e., *game app icons*) represented by the 22 individual adjective pairs and four variables related to icon success (1. overall evaluation of the icon, 2. willingness to click the icon, 3. willingness to download the imagined app and, 4. willingness to purchase the imagined app). In publication 4, MLRA was carried out to observe the effects of age, gender and times browsing app stores (per week) relating to perceptions of GUI element (i.e., *game app icon*) aesthetics. MLRA was also utilized to investigate the interaction terms of these independent variables, namely age x gender, age x time, gender x time and age x gender x time. The analyses were performed with the VISQUAL models as well as the 22 individual adjective pairs. The independent variables age and time were centered prior to the analyses (Aiken and West, 1991), and the interaction terms were created from the prior centered variables.

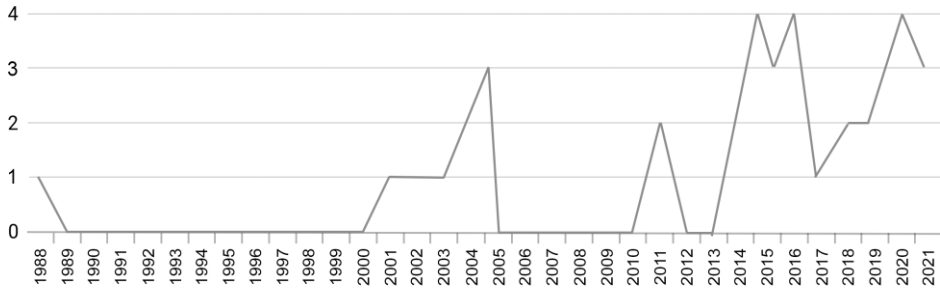
4 RESULTS

In order to fill the gaps identified by this research, and to discover findings as well as avenues for further research, four studies with various methodologies and a large dataset were carried out. Here, the results of these studies are presented by each publication. The findings of publication 1, a literature review, introduce the state-of-the-art of prior instruments evaluating visual aesthetics of graphical user interfaces. The results of publication 2 map out how VISQUAL, a psychometric instrument for measuring graphical user interface aesthetics, was composed and validated with further analyses. Furthermore, this section of results includes a guideline on how to use the instrument in future studies. Publication 3 investigated the key elements to successful graphical user interface graphics by user perception, measured by user evaluation and willingness to interact with the graphic. The findings underline overall high quality of the graphics with an illustrative example of icon material in the study with the highest score in user evaluations (Table 6). Publication 4 observed how demographic factors affect user perception of GUI aesthetics. The findings show what kinds of aesthetic perceptions graphical user interface elements should be brought to evoke, considering different target groups. Parts of this chapter have been previously published in peer-reviewed journal articles (Jylhä and Hamari 2019, 2020, 2021).

4.1 Publication 1

In publication 1, a systematic literature review was conducted to map out the state-of-the-art of instruments evaluating visual aesthetics of graphical user interfaces, i.e., how visual aesthetics of graphical user interfaces can be modelled, evaluated, and assessed. As presented in Figure 3, the review revealed an increase in relevant research. From 1988 to 2000 only one relevant study was published, whereas during the recent years of 2020–2021, eight articles on the topic have been published. A natural cause to this is the growing popularity of graphical user interfaces in our everyday lives. However, even then the field remains relatively untouched.

Figure 3. Articles published by year



Out of the 36 articles included in the review, most of the literature (55.6%) focused on predicting aesthetics of graphical user interfaces by geometry-related and image-related aspects in order to avoid human involvement in the process, e.g., balance, equilibrium, and symmetry. In addition to instruments that evaluate graphical assets by their features, aesthetic value of graphical user interfaces was measured by approaches by human perceptions (22.2%) with several different dimensions relating to e.g., quality, simplicity and aesthetic value. Recently, other methods (19.4%) have started to appear, that combine metrics and perceptions, and apply novel technologies, such as deep learning and eye-tracking. The majority of the approaches evaluate GUIs as a whole instead of separate GUI elements, and the context where the methods are developed is most commonly website evaluation. Overall, most of the studies (72.2%) were validated as opposed to non-validated studies. Regardless, the instruments remain scarce and scattered, which hinders cross-comparison and recommendation of methods to use in different contexts. Therefore, the extensive research agenda of publication 1 that explores multiple characteristics simultaneously with a wide perspective on the development of measurement instruments for graphical user interface aesthetics aimed to identify and bring light into these topical concepts. The results significantly narrow the gap in the current body of literature, however, more research is needed in order to accurately assign methods for various uses.

4.2 Publication 2

In the course of the research conducted and described in publication 2, the measurement instrument VISQUAL was developed, validated, and consequently

presented for perceptions of visual qualities of graphical user interfaces and/or singular interface elements that can be used in multiple ways in several contexts related to HCI, user interfaces, and their adaptation. The research is topical especially in the discourse of mobile platforms and the extremely competitive app markets, where visual design is considered one of the key factors of product success.

The instrument was developed in three consecutive stages. The initial measurement model of 22 items formed a five-factor structure in the EFA in the first stage of the development process. The factors were named to correspond 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 the second stage of development exposed concerns in the model, which were countered by item removal in the third stage. The adjusted model retained 15 (68%) items of the initial 22. This resulted in better validity and reliability producing more robust factors, thereby theoretically justifying this choice.

During the third stage, 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 similarly 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.

Publication 2 was a first-time evaluation and validation study for the measurement instrument VISQUAL. The instrument was developed for aiding research and design of aesthetically pleasing 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 merely at acceptable level. 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 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 as well as the adjusted model as presented in Table 11 is recommended in further assessment of the instrument. This means that all items marked ‘Yes’ should be used, however, we also recommend including the ‘Optional’ items when administering VISQUAL.

Table 11. Items in VISQUAL

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
Idleness/ Liveliness	Warm–Cool	Optional
	Slow–Fast	Yes
	Quiet–Loud	Yes
	Calm–Exciting	Yes
	Passive–Active	Yes
Normalness/	Old–Young	Optional
	Concrete–Abstract	Yes

Bizarreness	Realistic–Unrealistic Ordinary–Unique	Yes Optional
Complexity/ Simplicity	Complex–Simple Three-dimensional–Two- dimensional	Yes Yes

4.3 Publication 3

The objective of publication 3 was to observe how people’s aesthetic perceptions of GUI elements (i.e., *game app icons*) affect people’s willingness to interact with those elements, measured by 1) overall evaluation of the icon, 2) willingness to click the icon, 3) willingness to download the imagined app and, 4) willingness to purchase the app. The following results are based on qualitative assessment of the mean scores of different adjectives in icon ratings. The results can be applied to discussing best practices related to any graphical user interface elements, but naturally with caution as the external validity diminishes the more general the context in which the knowledge from the results is applied in.













First and foremost, the results unsurprisingly suggest that to increase consumer interaction in terms of app icon successfulness (i.e., overall evaluation, willingness to click an icon as well as download and purchase the imagined app behind the icon), the app icon should be perceived as high quality as indicated by the results where the following perceptions predicted app icon successfulness across the board: *beautiful, good, professional, and expensive*. All these adjectives can be associated with general high quality. If cursively investigating the icons that score high on these perceptions, they seem to share some of the following features: transparent parts on the outer layers, color gradients, shading and highlighting as well as high graphical fidelity.













Separately from perceptions related to high quality, *uniqueness* was another strong predictor of icon successfulness. Therefore, a successful app icon should be unique and memorable to stand out from the app store masses where there is a lot of icon material. If cursively investigating the icons that score high or low on the uniqueness–ordinariness continuum, they seem to share some of the following features: 1) icons rated as unique more commonly featured asymmetric and abstract shapes, use of various ends of the color spectrum as well as elements rarely encountered in daily life (e.g., a voodoo doll); 2) icons rated as ordinary broadly portrayed concrete, static shapes as well as objects typical to daily life (e.g., a house or a book).

Beyond all perceptions that predicted all other factors of icon successfulness, *sadness* and *fastness* weakly predicted willingness to purchase the imagined app behind the icon. If cursorily investigating the icons that score high on these factors, they seem to share some of the following features: 1) icons rated as sad were generally dominated by a desaturated or dark color palette (e.g., shades of grey or pure black), and they depicted elements that can be perceived as unpleasant; 2) icons rated as fast illustrated actions or objects that are typically associated with being rapid (e.g., a motorcycle or an airplane). A related observation is that icons with high scores of perceptions for things that are generally regarded as positive do not necessarily lead to higher icon successfulness. The indication that sadness predicts the willingness to purchase an app behind the icon is one example of this. Moreover, high overall evaluation score does not automatically lead to a high score in the willingness to click the icon, nor in the willingness to download or purchase the imagined app. Thus, it can be argued that app icons should incorporate more than one of the aforementioned implications in order to enhance the likelihood to consumer interaction.

Purely as illustrative examples, Table 12 introduces icons with the highest scores in overall evaluation of the icon design, the willingness to click the icon, as well as the willingness to download and purchase the imagined app. However, we wish to note that publication 3 did not investigate the relationship between icon features and successfulness per se. Therefore, any relationship between icon feature and success should be regarded as background data augmenting the focus of the study that was on the relationship between perception and successfulness.

Table 12. Top 6 icons with highest score (1 = lowest and 7 = highest)

#	Overall evaluation		Willingness to click		Willingness to download		Willingness to purchase	
	Icon	Mean	Icon	Mean	Icon	Mean	Icon	Mean
1		4.77		4.22		4.00		3.68
2		4.52		4.21		3.95		3.63
3		4.51		4.19		3.85		3.58

4		4.50		4.09		3.81		3.50
5		4.31		4.05		3.77		3.48
6		4.24		3.98		3.77		3.40

4.4 Publication 4

In the context of graphical user interface aesthetics, demographic factors have received minimal attention regardless of their relevance in design and development. Appealing graphical elements that cater to user needs are considered progressively important, as the way a graphic is visually represented can greatly contribute to the interaction. However, aesthetic perceptions are subjective and may differ by target group. Understanding variations in user perceptions may aid in design processes, therefore, demographic effects of age, gender and time using graphical user interfaces (i.e., *app stores*) relating to perceptions of GUI element (i.e., *game app icon*) aesthetics were investigated in publication 4.

The results indicate that, overall, demographic factors have relatively little effect on how icons are perceived. Nevertheless, these findings inform what kinds of aesthetic perceptions graphical user interface elements should be brought to evoke. This knowledge can then be adapted in establishing segmentation models for the design of adaptive user interfaces.

The findings show that younger users in general, as well as older women, tend to be more critical towards icon aesthetics. Thus, in order to visually appeal to the tastes of younger audiences and women, focusing on creating high quality designs (i.e., *high graphical fidelity*) is recommended, as the hedonic aspects need to be catered to across these demographic factors. Expectedly, time affects perceptions in that novice users perceived icons as more excellent than experienced users. Therefore, in order to visually appeal to more experienced users, designers may have to put in more effort and creativity. Overall, gender differences among younger users seem to be minimizing and therefore gender-neutral options could be considered in future design processes. However, the perceptions of icons change especially for younger

women in that icons are seen as more concrete and exciting over time. Hence, practitioners could benefit from integrating young female users to interfaces at an early stage to increase the aforementioned effects.

4.5 Overview of the results

The overarching objective in this dissertation was to investigate *how the visual aesthetics of graphical user interfaces can be modelled, evaluated, and assessed*. For this purpose, the paradigm was divided into the following research questions:

RQ1: What is the current state-of-the-art in the literature of measurement instruments of visual aesthetics for graphical user interfaces?

RQ2: What are the psychometric properties of VISQUAL, and what dimensions of aesthetic perception emerge (in the context of GUI elements)?

RQ3: How do perceived visual aesthetics predict the attractiveness of graphical user interface elements?

RQ4: Do demographic factors affect aesthetic perceptions of graphical user interface elements?

As described in the results for each publication, the separate research questions have been adequately answered. The results collected through the literature review (publication 1) clarify the current scattered measurement instruments by summarizing the measurement instruments by prior literature. The findings from the online survey and vignette experiment (publications 2, 3, and 4) produced a detailed description of the development of the measurement instrument VISQUAL, its psychometric properties and dimensions of aesthetic perception; surveying user perceptions in evaluations and willingness of interaction, perceived visual aesthetics were found to predict the attractiveness of GUI elements and a set of design guidelines were created based on the observations; and it was found that demographic factors minimally affect aesthetic perceptions, however, target groups are to be considered in design.

Overall, the studies presented here indicate that visual aesthetics, albeit subjective, follow tendencies and guidelines that are discoverable by a set of measurement

instruments and user perceptions. Furthermore, despite the complexity of modelling and evaluating the attractiveness of graphical user interfaces and elements, the tool developed in publication 2 (VISQUAL) succeeded in assessment, which is a major addition to the larger paradigm of the possibilities of design research. Considering this overview, the findings of the publications in this dissertation cover the overarching research objective to a satisfactory degree. The results have indeed paved the way for deeper investigation of modelling, evaluating, and assessing graphical user interface aesthetics. Distinct contributions include the implications on measurement instrument categorization enabling comparison and context-dependency of available tools for assessing the attractiveness of GUI elements (publication 1), a measurement instrument that was developed and validated in the course of this research (publication 2), as well as the collection of a large-scale dataset of user perceptions and the processing of this data into 1) confirming that aesthetics is an imperative part of general GUI satisfaction; 2) GUI element attractiveness can be predicted, and; 3) design guidelines can be modelled based on these analyses. However, the full potential of assessing and generating graphical user interface aesthetics is yet to be seized, which is discussed in the following.

5 DISCUSSION

This dissertation aimed to shed light on *how the visual aesthetics of graphical user interfaces can be modelled, evaluated, and assessed*, with particular focus on creating an overview of the field, while also examining user perceptions and emerging dimensions of aesthetics. Four publications were introduced as part of the dissertation in order to answer the research questions outlined in the introduction. As a summary, publication 1 described a systematic mapping of the state-of-the-art in the literature of measurement instruments of visual aesthetics for graphical user interfaces. Publication 2 continued in a similar thematic setting, observing what aesthetic features appear together in graphical icons, and consequently developing a psychometric scale that measures graphical user interface elements via individual user perceptions. Publication 3 then examined these user perceptions in order to find out how perceived visual aesthetics predict the attractiveness of graphical user interface elements. Finally, publication 4 examined demographic factors and their effects on aesthetic perceptions of graphical user interface elements.

Publication 1 succeeded in providing valuable information on the measurement instruments, their nature, similarities, and differences. In particular, a steady increase was found in evaluation methods that aim for assessment without human involvement, processing aesthetics by metrics and algorithms. Further, it was revealed that although validated instruments exist, they are scarce and scattered, which hinders cross-comparison. The extensive research agenda explored multiple characteristics simultaneously with a wide perspective on the development of measurement instruments for graphical user interface aesthetics.

Publication 2 presented a first-time evaluation and validation study for VISQUAL. The instrument was developed in the pursuit of aiding research and design of attractive interface elements, which has been lacking in the field of HCI. In this era of attention economy and everyday technology, 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 rather than more specific psychological traits. While aesthetic perception is subjective, this 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.

Publication 3 explored how user perceptions affect GUI element effectiveness from an aesthetic perspective, located in the domain of mobile game apps. The findings showed evidence of consensus that proves an empirical relationship on user perceptions and GUI element effectiveness. Revealing this phenomenon is a building block that has a potential for further theoretical implications around the topic. As such, publication 3 has laid the groundwork for future research that aims to understand user perceptions of GUI elements, and especially game app icons in graphical user interfaces and online storefronts.

Publication 4 showed that, overall, demographic factors have relatively little effect on user perceptions of GUI element aesthetics. However, experienced users, young audiences and women tend to be more critical towards aesthetics, in that they are more demanding in what is aesthetically pleasing. The general opinion was, however, that aesthetics matter. This concludes the overarching discourse that has been proven by all the publications in this dissertation, namely, that paying close attention to attractiveness is important in GUI design and a significant factor in determining the success of an interface.

Together these publications form a cohesive image on the key questions presented in this dissertation. While aesthetic appeal is subjective, the findings show evidence of consensus that proves an empirical relationship on user perceptions and the effectiveness of attractive GUI design. The current undertaking shows that technology adoption advances at a tremendous pace, which blurs the boundaries of aesthetics between people despite their age, gender and habits in daily life. Considering the changing cultural atmosphere, especially relating to gender and age in the domain of technology, insight into the topic is especially valuable. Therefore, this study has laid the groundwork that aims to understand perceptions of attractive graphical user interfaces, its current state-of-the-art and the many ways to investigate them.

5.1 Contributions

A shared trait to the publications in this dissertation is that because previous work has been lacking in this area, they introduce one of the first empirical studies dedicated to individual user perceptions and visual aesthetics located in the larger domain of graphical user interface research, and more specifically, the rapidly growing mobile apps and games markets, where minimal attention has been provided to how the visual attributes affect users. Therefore, a number of concrete contributions can be derived from this work.

A thorough overview of available measures was presented in this dissertation, which facilitates selecting tools for assessing GUI aesthetics and further studies on the topic. This can be seen as foundational work for the field as the measurement instruments have been previously scattered and thus some have been more discoverable than others, which implies a bias in the usage of these tools. With this contribution, the aim is to offer an objective outlook and possibility to select and compare measurement instruments and/or further develop them.

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 (Gullà et al., 2015). Prior research (e.g., Ngo et al., 2000; Ngo, 2001; Ngo et al., 2003; Vanderdonckt and Gillo, 1994; Zen and Vanderdonckt, 2014, 2016) 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. One of the main contributions of this research is an instrument developed and validated in publication 2, namely VISQUAL, 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. 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.

Following the lines of guiding practices, another main contribution that can be extracted from publications 2, 3 and 4 is the information on what kinds of aesthetic perceptions graphical user interface elements (i.e., icons) should be brought to evoke. As of yet, few guidelines exist on this topic, thus it adds significantly to the current body of literature. This knowledge can be adapted in establishing segmentation models for the design of customizable user interfaces for different target groups and use contexts. These design implications underline the importance of creating high quality designs (i.e., *transparent parts on the outer layers, color gradients, shading and highlighting as well as high graphical fidelity*) across all graphical elements and demographic groups present in this research. Another design implication is tied to creativity, also emphasized by prior research (Arend et al., 1987, Creusen and Schoormans, 2005, Creusen et al., 2010, Dewar, 1999, Goonetilleke et al., 2001, Huang et al., 2002), which is essential in this time where users scroll past thousands of images per day. With the increase of time, users will naturally adapt to aesthetics that essentially repeat similar patterns, which may lead to developing a critical eye towards graphical elements. This way the users establish a taste for GUI aesthetics over time, which might make users more selective. Therefore, a successful graphic element should be unique and memorable (i.e., *asymmetric, abstractly shaped*) to stand out from masses.

Concerning demographics, gender differences among younger users seem to be minimizing and therefore gender-neutral options could be considered in future design processes (Boiano et al., 2006, Morris et al., 2005). However, the perceptions of icons change especially for younger women in that icons are seen as more concrete and exciting over time. Hence, practitioners could benefit from integrating young female users to interfaces at an early stage to increase the aforementioned effects.

In a more detailed sense, the results in publication 3 exposed gaps in prior GUI element design theories that underline the significance of concreteness and abstractness (e.g. 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, McDougall et al., 2000, Moyes and Jordan, 1993, Rogers and Osborne, 1987), as well as complexity and simplicity (e.g. Choi and Lee, 2012, Goonetilleke et al., 2001, McDougall and Reppa, 2008, McDougall et al., 2013, McDougall et al., 2016) in aesthetic evaluation. The results contrast this notion as none of the

forementioned adjectives were statistically significant. This calls for more research on these particular aesthetic features.

5.2 Limitations

This dissertation presents several insights into the facets of visual aesthetics in the context of graphical user interfaces as well as their measurement and prediction, but as is natural to any research, some compromises have to be made. Hence, the limitations of this dissertation are outlined as follows.

Concerning methodology, the literature review in publication 1 was conducted via Scopus database. Although the query keywords and review process were meticulously constructed and applied, omission of relevant studies is possible due to the nature of systematic literature reviews. To alleviate this limitation, study selection and data extraction was performed with a rigorous protocol with a detailed description of the search and results.

The experiment data in publications 2, 3 and 4 was gathered via an online survey that has some shortcomings worth acknowledging. As the survey was a self-administered online task, there is a lack of control over respondent behavior and a potential for misinterpretation of the survey, posing a risk of common method bias (Podsakoff et al., 2003). Moreover, the survey was distributed on Finnish student organizations' mailing lists, thus the sample can be considered fairly homogenous. The majority of the respondents were from the same age group and from a similar cultural background, which could affect perceptions in the study. This is clearly visible in the results of publication 4, where the age representation focuses on younger age groups due to the limitations of the sample. Thus, the possible effects of age are not adequately shown in the results and should be substantially expanded with a broader sample in order to accurately answer the research questions regarding age in the publication.

As the sample is a nonprobability convenience sample, it is not necessarily representative of all users across all graphical user interfaces. Measures were taken to counter these issues, including e.g., randomization of items, survey piloting, and elimination of faulty responses. As the sample size is fairly large ($n = 569$), the generalizability of the dataset can be considered adequate. However, complementary methods with increased diversity would provide a more comprehensive overview on the topic.

A further limitation connected to the data collection method in publications 2, 3 and 4 is the use of a newly developed scale VISQUAL. While the instrument was composed by merging existing measures and those theorized by researchers but not previously tested, it showed promise regarding internal consistency. Nevertheless, there is a lack of clear construct validation, e.g., through correlation with related measurement instruments, experimental analyses, or a differentiation from constructs that are less related by definition. Also, problems remained with validity and reliability. While it can be speculated that the overall level of reliability and validity possible to be attained by attitudinal measurement instruments where data is based on subjective intercorrelations may not be satisfactory in general, additional confirmatory studies are required to further examine the validity of the measurement model. Additionally, using the same sample in both EFA and CFA bears the risk of sampling effects and overfitting, as the CFA will be repeating many of the relationships that were established through the EFA which decreases overall generalizability of the results. There could also be underlying effects of the particular sample that might not be found in a second sample (Bandalos and Finney, 2010; Fokkema and Greiff, 2017). Therefore, additional studies with more diverse datasets are needed in order to examine the psychometric validity of the measurement instrument. With regards to study material in publications 2, 3 and 4, one larger domain of iconography was selected as stimulus, namely game app icons. This decision was made to maximize internal validity, as 1) game app icons are internally a homogenous category of graphical elements in the sense that they all share the same size, same possible color space and thus eliminate unforeseen variability; 2) participants are usually familiar with icons and can therefore more effortlessly imagine seeing such icons in their normal life, facilitating the responses, 3) game icons exhibit perhaps more heterogeneity in possible styles compared with icons related to utilitarian software/apps, thus, game apps may offer greater external validity and/or generalizability across icons, 4) currently game apps represent a hugely timely phenomenon as game apps are clearly the most popular app category globally by several statistics (e.g., Statista, 2016; 2021; Newzoo, 2019).

In the experiment design, participants were uninformed about the purpose of the apps behind the icons in the experiment, as this could affect the results. Knowledge of the app may pose a risk of bias in user perceptions, thus the choice was made to eliminate possible confounding variables influencing the main objective of the study, namely aesthetics. This was further controlled by selecting new game app icons for the experiment that were not widely known. Furthermore, as is commonplace within the industry, actual data on app store usage was not available, thus the measurement

used in this study reports intended behavior with a vignette-style experiment setting. This may have an impact on the generalizability of the findings. Moreover, as the results consist of perceptions measured by quantitative means, the findings may be considered ambiguous with underlying biases.

5.3 Research avenues

Although graphical user interfaces are a significant part of our everyday lives, research in the field has remained surprisingly scarce. This merits further studies particularly due to the current rapid development of technology. To conclude, this dissertation offers several avenues for future research that have been discovered in the course of research work.

While the publications included in this dissertation narrowed down several gaps in prior literature, they revealed a lack of research in many areas, such as the aesthetic evaluation of specifically mobile GUIs, as well as comparison between desktop and mobile interface types in regards of aesthetics evaluation. In general, cross-comparison of measurement instruments that evaluate visual aesthetics of GUIs is challenging due to inconsistent validation and scattered tools, which requires further systematic mapping. Fortunately, the need for systematic mapping has been recognized in the field (Lima and Gresse von Wangenheim 2022), and it is expected that more publications will be seen in the near future where the various measurement instruments are further classified and organized. Additionally, measurement instruments addressing diverse target groups in relation to e.g., accessibility, color blindness, and various demographic factors across different cultures are missing. As adaptive user interfaces, interface customization as well as context-aware environments become increasingly common, and design patterns and models are explored for adaptive applications (Bouzit et al., 2017; Braham et al., 2022), further research would provide insight into the options to be designed and offered in the future.

The limitations discussed the methodology, experiment design and stimulus material employed in this research. To address the concerns of measuring perceptions by quantitative means and possible ambiguity in the results, acquiring a qualitative approach would be beneficial in order to gain a deeper understanding of the topic in further studies. The selection criteria for study material poses a possibility for conducting future research on other GUI element types for comparative results.

Relating to the validity and reliability of the new measurement instrument VISQUAL, a number of directions for future work could be taken in order to address the issues mentioned in this dissertation. Firstly, as the initial model contained gaps that were addressed in a post-hoc revision, this moved the investigation out of a confirmatory analytic framework. Therefore, a replication study is recommended to define the properties of the measurement model. Secondly, measuring user perceptions can be seen as an adequate approach for user modelling, but a complementary measurement model that investigates personality dimensions (i.e., attitudes, behavioral tendencies, technology acceptance, aesthetics preferences) could be developed. The exploration of the potential for psychophysiological measures to offer deeper understanding of GUI design in both user experience and aesthetics of HCI has already commenced (e.g., Maia and Furtado 2018; 2019; Pentus et al. 2014; Rui and Gu 2021). Following this direction would link individual user perceptions measured by VISQUAL with personality traits, which could then be used to determine further recommendations on adaptation and customization. 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. Lastly, VISQUAL was validated by measuring visual qualities of single GUI elements, 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.

6 CONCLUSION AND FUTURE AGENDA

At the time of this dissertation, people are embracing an increasingly mobile lifestyle facilitated through various user interfaces. The pandemic has taught us new ways of interaction, anywhere and anytime. Through rigorous analyses, reviews, and discussions, the four publications in the dissertation depict and further develop relevant theories, methods, and guidelines for the research and production of graphical user interface aesthetics. The results reflect new patterns in our everyday life concerning human-computer interaction, and the evolution will surely continue. Based on the discourse in this dissertation, a three-way future agenda was identified around the domains of concept, theme, and methodology.

Firstly, as for conceptual agenda, the current corpus on user interface aesthetics and methods concentrates on geometry-related and image-related aspects, usually with fabricated settings with geometrical figures, which does not necessarily replicate a real setting. These constructs can be considered intuitive starting points for research, however, new mobile habits will accelerate the development of technology, which pushes the need for maturity in research as well. Therefore, experimenting with settings that we currently use on a daily basis is needed to start figuring out the next stage of “what could be” in terms of user interfaces, their attractiveness, and how we utilize them. Graphical user interfaces are, in their essence, *technologies that facilitate the reality*. While we use them on a daily basis, the development of these technologies is still in its infancy. User interfaces are multi-modal virtual products that are in a large sense free from constraints. Regardless, use experience right now can be considered very limited in terms of visual, aural, oral, haptic, and movement possibilities. For example, advances in speech recognition and natural language processing have the potential to make user interfaces more intuitive and effective than ever, transcending language barriers and cultural differences. Currently, we are bound to this limited experience with which we carry out our daily tasks, sometimes with success, sometimes with frustration. The concept of user interface aesthetics can be used as a beginning to explore novel ways of interaction, but to reach its full potential, a larger epistemological mindset has to be realized that takes into account what is beyond the graphical user interface that is known to us at present, and how their look and feel can be brought to a whole new level.

Secondly, relating to thematic agenda, the experiences of different users should be investigated further. An efficient user interface provides high fidelity, aesthetically pleasing graphics, and is intuitive to use for people with different technology skills and cultural backgrounds. Nevertheless, often there is a disconnect between the interface and the person who uses it, resulting in a negative experience. While this is more evident with older age groups due to the so-called digital divide, younger age groups tend to be more critical towards the look and feel of user interfaces in general. The discussion on how to increase the appeal of user interfaces from the perspective of efficiency is still insufficient. For example, it should be investigated when and where, in what circumstances, does different UIs prove to be effective and efficient, and for whom. As such, personality traits and individual characteristics influencing perceptions of user interface look and feel should also be examined with a user's values and attitudes included. Better personalized, customized, and adaptive systems have shown to lead to more positive, long-term user experiences (Debevc et al., 1996; Hartmann et al., 2007; Sarsam and Al-Samarraie, 2018, thus attribute studies can help researchers and practitioners design interfaces that live up to the expectations.

Finally, concerning methodological agenda, future studies can carry out comparative experiments between user perceptions of the designs of different user interfaces and UI elements, as there is a lack of knowledge on clear implications or guidelines for the various user interface platforms (e.g., mobile and desktop) and UI elements (e.g., fonts, icons, and menus). Research has focused on a relatively small area of UI and element types, which hinders the reliability and generalizability of the results. More diversity is needed to acquire a deeper understanding on user needs. Future research can also benefit from the development of eye-tracking tools and deep learning approaches, as they have proven to be reliable in studying user perceptions related to the attractiveness and effectiveness of graphical user interface designs (Dou et al., 2019; Gu et al., 2020; Khani et al., 2016; Liu and Yiang 2021; Pappas et al., 2020; Wu et al., 2011; Wu et al., 2016; Xing et al., 2021). These approaches draw from the evaluation methods based on graphical features as well as human perceptions with massively large-scale datasets, which increases the validity and accuracy of results and reduces time used for experiments. This way, the scope of samples can be expanded in terms of quantity, age, gender, income, and nationality. Naturally, multiple high-quality data sources are needed with a mass of different UI layouts and graphical elements of varying quality and features, as well as user perceptions on the attractiveness of these graphics. To improve accuracy and bias, future studies can be enriched with psychophysiological that provide understanding of user behavior on

an implicit level. Such measures aside from eye-tracking include e.g., body movement trackers, blood pressure monitors, and electroencephalograms (EEG). This methodology is already utilized in a plethora of studies across different fields, but its use is still scarce relating to graphical user interface aesthetics. The decision whether an object is attractive or not is formed in the subconscious mind within seconds of our exposure to it, and sometimes these reactions may be lost in an experiment setting with interviews or surveys for example. Therefore, an all-encompassing research methodology with the measurement of emotional arousal levels could be ground-breaking and fundamental to the evolution of graphical user interface design and the multi-modal experience.

REFERENCES

- Ahmed S.U., Al Mahmud A. and Bergaust K. (2009), “Aesthetics in human-computer interaction: Views and reviews”, Jacko, J.A. (Ed.s.), *Human-Computer Interaction, New Trends, HCI 2009*, Lecture Notes in Computer Science, Vol. 5610, Springer, Berlin.
- Aiken, L.S. and West, S.G. (1991), *Multiple regression: Testing and interpreting interactions*, Sage, London.
- Alemerien, K. and Magel, K. (2015), “SLC: A visual cohesion metric to predict the usability of graphical user interfaces”, paper presented at the 30th Annual ACM Symposium on Applied Computing, Salamanca, Spain, pp.1526–1533. doi: 10.1145/2695664.2695791
- App Annie (2021), “2021: The year the world is set to spend \$135 billion dollars – In mobile apps and games in new record”, available at: <https://www.appannie.com/en/insights/market-data/2021-end-year-mobile-apps-recap> (accessed January 3, 2022).
- Arend, U., Muthig, K.P. and Wandmacher, J. (1987), “Evidence for global superiority in menu selection by icons”, *Behaviour & Information Technology*, Vol. 6 No. 4, pp.411–426. doi: 10.1080/01449298708901853
- Ares, G., Piqueras-Fiszman, B., Varela, P., Marco, R.M., López, A.M. and Fiszman, S. (2011), “Food labels: Do consumers perceive what semiotics want to convey?” *Food Quality and Preference*, Vol. 22 No. 7, pp.689–698. doi: 10.1016/j.foodqual.2011.05.006
- Bandalos, D. L. and Finney, S. J. (2010), “Factor analysis. Exploratory and confirmatory”, Hancock, G. R. and Mueller, R. O. (Eds.), *The reviewer’s guide to quantitative methods in the social science*, pp.93–114. New York: Routledge.
- Bessghaier, N., Soui, M., Kolski, C. and Chouchane, M. (2021), “On the detection of structural aesthetic defects of android mobile user interfaces with a metrics-based tool”, *ACM Transactions on Interactive Intelligent Systems*, Vol. 11 No. 1, pp.1–27. doi: 10.1145/3410468
- Blijlevens, J., Thurgood, C., Hekkert, P., Chen, L.-L., Leder, H., and Whitfield, T. W. A. (2017), “The Aesthetic Pleasure in Design Scale: The development of a scale to measure aesthetic pleasure for designed artifacts”, *Psychology of Aesthetics, Creativity, and the Arts*, Vol. 11 No. 1, pp.86–98. doi: 10.1037/aca0000098
- Blankenberger, S. and Hahn, K. (1991), “Effects of icon design on human-computer interaction”, *International Journal of Man-Machine Studies*, Vol. 35 No. 3, pp.363–377. doi: 10.1016/S0020-7373(05)80133-6
- Boiano, S., Borda, A., Bowen, J., Faulkner, X., Gaia, G. and Mcdaid, S. (2006), “Gender issues in HCI design for web access”, Kurniawan, S. and Zaphiris, P. (Ed.s.), *Advances*

in universal web design and evaluation: Research, trends and opportunities, IGI Global, London, pp.116–153.

- Bouzit, S., Calvary, G., Coutaz, J., Chêne, D., Petit, E. and Vanderdonck, J. (2017), “The PDA-LPA design space for user interface adaptation”, paper presented at the 11th International Conference on Research Challenges in Information Science (RCIS). Brighton, UK. doi: 10.1109/RCIS.2017.7956559
- Braha, B., Buendía, F., Khemaja, M. and Gargouri, F. (2022), “User interface design patterns and ontology models for adaptive mobile applications”, *Personal and Ubiquitous Computing*, Vol 26, pp.1395–1411.
- Brown, T.A. (2015), *Confirmatory factor analysis for applied research*. Guilford Publications, New York.
- Burgers, C., Eden, A., de Jong, R. and Buningh, S. (2016), “Rousing reviews and instigative images: The impact of online reviews and visual design characteristics on app downloads”, *Mobile Media & Communication*, Vol. 4 No. 3, pp.327–346. doi: 10.1080/15213269.2016.1182030
- Chartboost (2015), “Power-Up Report – July 2015”, available at: <https://chartboost.s3.amazonaws.com/blog/power-up-report-july-2015-building-an-empire-mobile-strategy-games.pdf> (accessed January 3, 2022).
- Chen, CC. (2015), “User recognition and preference of app icon stylization design on the smartphone”, Stephanidis, C. (Eds.), *HCI International 2015 - Posters' Extended Abstracts*, Vol 529. Springer, Cham. doi: 10.1007/978-3-319-21383-5_2
- Cho, H.-S. and Lee, J. (2005), “Development of a macroscopic model on recent fashion trends on the basis of consumer emotion”, *International Journal of Consumer Studies*, Vol. 29 No. 1, pp.17–33. doi:10.1111/j.1470-6431.2005.00370.x
- Choi, J. H. and Lee, H.-J. (2012), “Facets of simplicity for the smartphone interface: A structural model”, *International Journal of Human-Computer Studies*, Vol. 70 No. 2, pp.129–142. doi: 10.1016/j.ijhcs.2011.09.002
- Contributor Roles Taxonomy (CRediT) (2015), “CRediT author statement”, available at: <https://www.elsevier.com/authors/policies-and-guidelines/credit-author-statement> (accessed December 30, 2021).
- Creswell, J.W. and Clark, V.L.P. (2011), *Designing and conducting mixed methods research*, SAGE Publications.
- Creusen, M.E.H. and Schoormans, J.P.L. (2005), “The different roles of product appearance in consumer choice”, *Journal of Product Innovation Management*, Vol. 22 No. 1, pp.63–81. doi: 10.1111/j.0737-6782.2005.00103.x
- Creusen, M.E.H., Veryzer, R.W. and Schoormans, J.P.L. (2010), “Product value importance and consumer preference for visual complexity and symmetry”, *European Journal of Marketing*, Vol. 44 No. 9/10, pp.1437–1452. doi: 10.1108/03090561011062916
- Creusen, M.E.H. (2010), “The importance of product aspects in choice: The influence of demographic characteristics”, *Journal of Consumer Marketing*, Vol. 27 No. 1, pp.26–34. doi: 10.1108/07363761011012921

- Cyr, D. (2009), "Gender and website design across cultures", paper presented at the 17th European Conference on Information Systems, Verona, Italy, pp.279–291.
- Debevc, M., Meyer, B., Donlagic, D. and Svecko, R. (1996), "Design and evaluation of an adaptive icon toolbar", *User Modeling and User-Adapted Interaction*, Vol. 6 No. 1, pp.1–21. doi: 10.1007/BF00126652
- Dewar, R. (1999), "Design and evaluation of public information symbols", Zwaga, H. J. G., Boersema, T. and Hoonhout, H. C. M. (Ed.s.), *Visual Information for Everyday Use*. Taylor & Francis, London, pp.285–303.
- Dou, Q., Zheng, S., Sun, T. and Heng, P.A. (2018), "Webthetics: Quantifying webpage aesthetics with deep learning", *International Journal of Human-Computer Studies*, Vol. 124, pp.56–66. doi: 10.1016/j.ijhcs.2018.11.006
- Dutton, D. (2009), *The art instinct*, New York, NY: Oxford University Press.
- Fenko, A., Schifferstein, H. N. J. and Hekkert, P. (2010), "Shifts in sensory dominance between various stages of user-product interactions", *Applied Ergonomics*, Vol. 41 No. 1, pp.34–40. doi: 10.1016/j.apergo.2009.03.007
- Fokkema, M. and Greiff, S. (2017), "How performing PCA and CFA on the same data equals trouble: Overfitting in the assessment of internal structure and some editorial thoughts on it", *European Journal of Psychological Assessment*, Vol. 33 No. 6, pp.399–402. doi: 10.1027/1015-5759/a000460
- Fornell, C. and Larcker, D.F. (1981), "Evaluating structural equation models with unobservable variables and measurement error", *Journal of Marketing Research*, Vol. 18, pp.39–50. doi: 10.2307/3151312
- Gait, J. (1985), "An aspect of aesthetics in human-computer communications: Pretty windows", *IEEE Transactions on Software Engineering*, Vol. 11 No. 8, pp.714–717. doi: 10.1109/TSE.1985.232520
- García, M., Badre, A.N. and Stasko, J.T. (1994), "Development and validation of icons varying in their abstractness", *Interacting with Computers*, Vol. 6, pp.191–211. doi: 10.1016/0953-5438(94)90024-8
- Genuine (2013), "Gender-inclusive user interface guidelines", available at: <http://genuine.ict.tuwien.ac.at/unterlagen/Guidelines.pdf> (accessed January 2, 2022).
- Gittins, D. (1986), "Icon-based human-computer interaction", *International Journal of Man-Machine Studies*, Vol. 24, pp.519–543. doi: 10.1016/S0020-7373(86)80007-4
- Goonetilleke, R.S., Shih, H.M., On, H.K. and Fritsch, J. (2001), "Effects of training and representational characteristics in icon design", *International Journal of Human-Computer Studies*, Vol. 55 No. 5, pp.741–760. doi: 10.1006/ijhc.2001.0501
- Gu, Z., Jin, C., Dong, Z. and Chang, D. (2020), "Predicting webpage aesthetics with heatmap entropy", *Behaviour and Information Technology*, Vol. 40 No. 7, pp.676–690. doi: 10.1080/0144929X.2020.1717626

- Gullà F., Ceccacci S., Germani M., Cavalieri L. (2015), “Design adaptable and adaptive user interfaces: A method to manage the information”, Andò B., Siciliano P., Marletta V. and Monteriù A. (Ed.s.), *Ambient Assisted Living, Biosystems & Biorobotics*, vol 11. Springer, Cham, pp. 47–58.
- Hartmann, J., Sutcliffe, A. and De Angeli, A. (2007a), “Towards a theory of user judgment of aesthetics and user interface quality”, *ACM Transactions on Computer-Human Interaction*, Vol. 15 No. 4. doi: 10.1145/1460355.1460357
- Hartmann, J., Sutcliffe, A. and De Angeli, A. (2007b), “Investigating attractiveness in web user interfaces”, paper presented at the 25th Annual SIGCHI Conference on Human Factors in Computing Systems, San Jose, USA, pp.387–396.
- Hartmann, J., De Angeli, A. and Sutcliffe, A. (2008), “Framing the user experience: Information biases on website quality judgement”, paper presented at the 26th Annual SIGCHI Conference on Human Factors in Computing Systems, Florence, Italy, pp.855–864.
- Hassenzahl, M., Burmester, M. and Koller, F. (2003), “AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität [AttracDiff: A questionnaire to measure perceived hedonic and pragmatic quality]”, Ziegler, J. and Szwillus, G. (Ed.s.), *Mensch&Computer 2003 – Interaktion in Bewegung*, B. G. Teubner, Stuttgart, pp.187–196.
- Hassenzahl, M. (2004), “The interplay of beauty, goodness, and usability in interactive products”, *Human–Computer Interaction*, Vol. 19 No. 4, pp.319–349. doi: 10.1207/s15327051hci1904_2
- Henry, P. (2002), “Systematic variation in purchase orientations across social classes”, *Journal of Consumer Marketing*, Vol. 19 No. 5, pp.424–438. doi: 10.1108/07363760210437641
- Hermida, R. (2015), “The problem of allowing correlated errors in structural equation modeling: concerns and considerations”, *Computational Methods in Social Sciences*, Vol. 3 No. 1, pp.5–17.
- Hou, K.-C. and Ho, C.-H. (2013), “A preliminary study on aesthetic of apps icon design”, paper presented at the 5th International Congress of International Association of Societies of Design Research, Tokyo, Japan, pp.3845–2856.
- Horton, W. (1994), *The icon book: Visual symbols for computing systems and documentation*, John Wiley & Sons, New York.
- Horton, W. (1996), “Designing icons and visual symbols”, paper presented at the CHI 96 Conference on Human Factors in Computing Systems, Vancouver, Canada, pp.371–372. doi: 10.1145/257089.257378
- Huang, S. (2013), “Usability and GUI Design and Principles”, available at: https://www.uio.no/studier/emner/matnat/ifi/INF5120/v13/undervisningsmateriale/f04-2013_0402-3-designguidelines_additional_info.pdf (accessed January 2, 2022).

- Huang, S.-M., Shieh, K.-K. and Chi, C.-F. (2002), “Factors affecting the design of computer icons”, *International Journal of Industrial Ergonomics*, Vol. 29 No. 4, pp.211–218. doi: 10.1016/S0169-8141(01)00064-6
- Isherwood, S.J., McDougall, S.J.P. and Curry, M.B. (2007), “Icon identification in context: The changing role of icon characteristics with user experience”, *Human Factors*, Vol. 49 No. 3, pp.465–476. doi: 10.1518/001872007X200102
- Jankowski, J., Bródka, P. and Hamari, J. (2016), “A picture is worth a thousand words: an empirical study on the influence of content visibility on diffusion processes within a virtual world”, *Behaviour & Information Technology*, Vol. 35 No. 11, pp.926–945.
- Jankowski, J., Hamari, J. and Watrobski, J. (2019), “A gradual approach for maximising user conversion without compromising experience with high visual intensity website elements”, *Internet Research*, Vol. 29 No. 1, pp.194–217.
- Jennings, M. (2000), “Theory and models for creating engaging and immersive ecommerce websites”, paper presented at the 2000 ACM SIGCPR Conference on Computer Personnel Research, New York, USA, pp.77–85. doi: 10.1145/333334.333358
- Johnson, J. and Finn, K. (2017), *Designing user interfaces for an aging population: Towards universal design*, Morgan Kaufmann, Amsterdam.
- Jordan, P.W. (1998), “Human factors for pleasure in product use”, *Applied Ergonomics*, Vol. 29 No. 1, pp.25–33. doi: 10.1016/S0003-6870(97)00022-7
- Jylhä, H. and Hamari, J. (2021), “Demographic factors have little effect on aesthetic perceptions of icons: a study of mobile game icons”, *Internet Research*, Ahead-of-print. doi: 10.1108/INTR-07-2020-0368
- Jylhä, H. and Hamari, J. (2020), “Development of measurement instrument for visual qualities of graphical user interface elements (VISQUAL): A test in the context of mobile game icons”, *User Modeling and User-Adapted Interaction*, Vol. 30 No. 5, pp.949–982. doi: 10.1007/s11257-020-09263-7
- Jylhä, H. and Hamari, J. (2019), “An icon that everyone wants to click: How perceived aesthetic qualities predict app icon successfulness”, *International Journal of Human-Computer Studies*, Vol. 130, pp.73–85. doi: 10.1016/j.ijhcs.2019.04.004
- Kaplan, D. (1990), “Evaluating and modifying covariance structure models: A review and recommendation” *Multivariate Behavioral Research*, Vol. 25 No. 2, pp.137–155. doi: 10.1207/s15327906mbr2502_1
- Kenny, D.A. (2020), “Measuring Model Fit”, available at: <http://davidakenny.net/cm/fit.htm> (accessed January 3, 2022).
- Khani, M., Mazinani, M., Fayyaz, M. and Mojtaba, H. (2016), “A novel approach for website aesthetic evaluation based on convolutional neural networks”, paper presented at the Second International Conference on Web Research (ICWR), IEEE, Tehran, pp.48–53. doi: 10.1109/ICWR.2016.7498445.
- Kline, R.B. (2011), *Principles and practice of structural equation modelling*, Guilford Press, New York.

- KnowItAll Ninja (2016), “Exploring user interface design principles and project planning techniques”, available at: <https://www.pearsonschoolsandfecolleges.co.uk/AssetsLibrary/SECTORS/FurtherEducationColleges/SUBJECT/dit/dit-student-book/btec-techaward-dit-sb-9781292208374.pdf> (accessed January 2, 2022).
- Kohonen, I., Kuula-Luumi, A. and Spoof, S. K. (2019), *The ethical principles of research with human participants and ethical review in the human sciences in Finland*, Helsinki: Finnish National Board on Research Integrity TENK.
- Kurosu, M. and Kashimura, K. (1995), “Apparent usability vs. inherent usability”, paper presented at the CHI 95 Conference Companion on Human Factors in Computing Systems, New York, USA, pp.292–293. doi: 10.1145/223355.223680
- Labrecque, L. and Milne, G. (2011), “Exciting red and competent blue: The importance of color in marketing”, *Journal of the Academy of Marketing Science*, Vol. 40 No. 5, pp.711–727. doi: 10.1007/s11747-010-0245-y
- Lavie, T. and Tractinsky, N. (2004), “Assessing dimensions of perceived visual aesthetics of web sites”, *International Journal of Human-Computer Studies*, Vol. 60 No. 3, pp.269–298. doi: 10.1016/j.ijhcs.2003.09.002
- Leder, H., Belke, B., Oeberst, A. and Augustin, D. (2004), “A model of aesthetic appreciation and aesthetic judgments”, *British Journal of Psychology*, Vol. 95, pp.489–508. doi: 10.1348/0007126042369811
- Lee, S.H. and Boling, E. (1999) “Screen design guidelines for motivation in interactive multimedia instruction: A survey and framework for designers”, *Educational Technology archive*, Vol. 39, pp.19–26.
- Lee, S. and Koubek, R.J. (2011), “The impact of cognitive style on user preference based on usability and aesthetics for computer-based systems”, *International Journal of Human-Computer Studies*, Vol. 27 No. 11, pp.1083–1114. doi: 10.1080/10447318.2011.555320
- Lima, de Souza A. L. and Gresse von Wangenheim, C. (2022), “Assessing the visual esthetics of user interfaces: a ten-year systematic mapping”, *International Journal of Human-Computer Interaction*, Vol. 38 No. 2, pp.144–164.
- Lin, C.-H. and Chen, M. (2018), “The icon matters: How design instability affects download intention of mobile apps under prevention and promotion motivations”, *Electronic Commerce Research*, Vol. 1. doi: 10.1007/s10660-018-9297-8
- Lin, C.-L. and Yeh, J.-T. (2010), “Marketing aesthetics on the web: Personal attributes and visual communication effects”, paper presented at the 5th IEEE International Conference on Management of Innovation & Technology, IEEE, Singapore, pp.1083-1088.
- Linux Information Project (2004), “GUI Definition”, available at: <http://www.linfo.org/gui.html> (accessed December 5, 2021).
- Liu, X. and Jiang, Y. (2020), “Aesthetic assessment of website design based on multimodal fusion”, *Future Generation Computer Systems*, Vol. 117, pp.433–438. doi: 10.1016/j.future.2020.12.014

- Lodding, K.N. (1983), "Iconic interfacing", *IEEE Computer Graphics and Applications*, Vol. 3, pp.11–20. doi: 10.1109/MCG.1983.262982
- MacCallum, R. (1986), "Specification searches in covariance structure modeling", *Psychological Bulletin*, Vol. 100 No. 1, pp.107–120. doi: 10.1037/0033-2909.100.1.107
- MacKenzie, S.B., Podsakoff, P.M. and Podsakoff, N.P. (2011), "Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques", *MIS Quarterly*, Vol. 35 No. 2, pp.293–334. doi: 10.2307/23044045
- Matsunaga, M. (2010), "How to factor-analyze your data right: Do's, don'ts, and how-to's", *International Journal of Psychological Research*, Vol. 3 No. 1, pp.97–110. doi: 10.21500/20112084.854
- Maia, C.L.B. and Furtado, E. S. (2018), "Using psychophysiological measures to estimate dimensions of emotion in hedonic experiences", *Computers & Electrical Engineering*, Vol. 71, pp.431–439. doi: 10.1016/j.compeleceng.2018.07.048
- Maia, C.L.B. and Furtado, E. S. (2019), "An approach to analyze user's emotion in hci experiments using psychophysiological measures", *IEEE Access*, Vol. 7. doi: 10.1109/ACCESS.2019.2904977
- Maity, R., Uttav, A., Gourav, V. and Bhattacharya, S. (2015), "A non-linear regression model to predict aesthetic ratings of on-screen images", paper presented at the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction, OZCHI 2015, Parkville, Australia, pp.44–52. doi: 10.1145/2838739.2838743
- Maity, R., Madrosiya, A. and Bhattacharya, S. (2016), "A computational model to predict aesthetic quality of text elements of GUI", *Procedia Computer Science*, Vol. 84, pp.152–159. doi: 10.1016/j.procs.2016.04.081
- Maity, R. and Bhattacharya, S. (2019), "Is My Interface Beautiful? – A Computational Model-Based Approach", *IEEE Transactions on Computational Social Systems*, Vol. 6 No. 1, pp.149–161. doi: 10.1109/TCSS.2019.2891126
- Mbenza, P. and Burny, N. (2021), "Computing aesthetics of concrete user interfaces", paper presented at the 12th ACM SIGCHI Symposium on Engineering Interactive Computing Systems, ACM, pp. 1–8.
- McDougall, S.J.P., Curry, M.B. and de Bruijn, O. (1998), "Understanding what makes icons effective: How subjective ratings can inform design", Hanson, M. (Eds.), *Contemporary Ergonomics*. Taylor & Francis, London, pp.285–289.
- McDougall, S.J.P., Curry, M.B. and de Bruijn, O. (1999), "Measuring symbol and icon characteristics: Norms for concreteness, complexity, meaningfulness, familiarity, and semantic distance for 239 symbols", *Behavior Research Methods, Instruments, & Computers*, Vol. 31 No. 3, pp.487–519. doi: 10.3758/BF03200730
- McDougall, S.J.P., de Bruijn, O. and Curry, M.B. (2000), "Exploring the effects of icon characteristics on user performance: The role of icon concreteness, complexity, and distinctiveness", *Journal of Experimental Psychology: Applied*, Vol. 6 No. 4, pp.291–306. doi: 10.1037/1076-898X.6.4.291

- McDougall, S.J.P. and Reppa, I. (2008), “Why do I like it? The relationships between icon characteristics, user performance and aesthetic appeal”, paper presented at the Human Factors and Ergonomics Society 52nd Annual Meeting, New York, USA, pp.1257–1261. doi: 10.1177/154193120805201822
- McDougall S.J.P. and Reppa I. (2013), “Ease of icon processing can predict icon appeal”, paper presented at the 15th International Conference on Human-Computer Interaction, Las Vegas, USA, pp.575–584.
- McDougall, S.J.P., Reppa, I., Kulik, J. and Taylor, A. (2016), “What makes icons appealing? The role of processing fluency in predicting icon appeal in different task contexts”, *Applied Ergonomics*, Vol. 55, pp.156–172. doi: 10.1016/j.apergo.2016.02.006
- Meier, B.J. (1988), “ACE: A color expert system for user interface design”, paper presented at the 1st Annual ACM SIGGRAPH Symposium on User Interface Software and Technology, ACM, pp.117–128.
- Miniukovich, A. and De Angeli, A. (2014a), “Quantification of interface visual complexity”, paper presented at the Workshop on Advanced Visual Interfaces AVI, ACM, New York, pp.153–160. doi: 10.1145/2598153.2598173.
- Miniukovich, A. and De Angeli, A. (2014b), “Visual impressions of mobile app interfaces”, paper presented at the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational, Helsinki, Finland, pp.31-40. doi: 10.1145/2639189.2641219
- Miniukovich, A. and De Angeli, A. (2015a), “Computation of interface aesthetics”, paper presented at the 33rd Annual Conference on Human Factors in Computing Systems, ACM, pp.1163–1172.
- Miniukovich, A. and De Angeli, A. (2015b), “Visual diversity and user interface quality”, paper presented at the 2015 British HCI Conference, ACM, New York, pp.101–109. doi: 10.1145/2783446.2783580
- Miniukovich, A., Sulpizio, S. and De Angeli, A. (2018), “Visual complexity of graphical user interfaces”, paper presented at the 2018 International Conference on Advanced Visual Interfaces, ACM, New York, pp.1–9. doi: 10.1145/3206505.3206549
- Morris, M.G., Venkatesh, V. and Ackerman, P.L. (2005), “Gender and age differences in employee decisions about new technology: An extension to the theory of planned behavior”, *IEEE Transactions on Engineering Management*, Vol. 52 No. 1, pp.69–84.
- Moshagen, M. and Thielsch, M. (2010), “Facets of visual aesthetics”, *International Journal of Human-Computer Studies*, Vol. 68 No. 4, pp.689–709. doi: 10.1016/j.ijhcs.2010.05.006
- Moshagen, M. and Thielsch, M. (2013), “A short version of the visual aesthetics of websites inventory”, *Behaviour & Information Technology*, Vol. 32 No. 12, pp.1305–1311. doi: 10.1080/0144929X.2012.694910
- Moyes, J. and Jordan, P.W. (1993), “Icon design and its effect on guessability, learnability, and experienced user performance”, Alty, J.D., Diaper, D. and Gust, S. (Ed.s.), *People and computers VIII*. Cambridge University Society, Cambridge, pp.49–59.
- Newzoo (2019), “The Global Games Market Will Generate \$152.1 Billion in 2019 as the U.S. Overtakes China as the Biggest Market”, <https://newzoo.com/insights/articles/the->

global-games-market-will-generate-152-1-billion-in-2019-as-the-u-s-overtakes-china-as-the-biggest-market/ (accessed January 2, 2022).

- Ngo, D.C.L., Samsudin, A. and Abdullah, R. (2000), "Aesthetic measures for assessing graphic screens", *Journal of Information Science and Engineering*, Vol. 16 No. 1, pp.97–116.
- Ngo, D. C. L., and Byrne, J. G. (2001), "Application of an aesthetic evaluation model to data entry screens", *Computers in Human Behavior*, Vol. 17 No. 2, pp.149–185.
- Norman, D.A. (2004), *Emotional design: Why we love (or hate) everyday things*, Basic Books, New York.
- Nunnally, J.C. and Bernstein, I. (1994), *Psychological theory*, McGraw-Hill, New York.
- Orth, U.R. and Malkewitz, K. (2008), "Holistic package design and consumer brand impressions", *J. Mark.* 72, 64–81. doi: 10.1509/jmkg.72.3.64
- Orth, U.R. and Malkewitz, K. (2009), "Good from far but far from good: The effects of visual fluency on impressions of package design", *Adv. Consum. Res.* 36, 211–212.
- Overby, E. and Sabyasachi, M. (2014), "Physical and electronic wholesale markets: An empirical analysis of product sorting and market function", *Journal of Management Information Systems*, Vol. 31 No. 2, pp.11–46. doi: 10.2753/MIS0742-1222310202
- Oyibo, K. and Vassileva, J. (2017), "The interplay of aesthetics, usability and credibility in mobile website design and the moderation effect of gender", *Journal on Interactive Systems*, Vol. 8 No. 2, pp.4–19.
- Oyibo, K., Adaji, I. and Vassileva, J. (2018), "The effect of age and information design on the perception of visual aesthetic", paper presented at the British Human Computer Interaction Workshop, Belfast, UK. doi: 10.14236/ewic/HCI2018.208
- Pappas, I.O., Mikalef, P., Giannakos, M.N. and Kourouthanassis, P.E. (2019), "Explaining user experience in mobile gaming applications: an fsQCA approach", *Internet Research*, Vol. 29 No. 2, pp.293–314. doi: 10.1108/IntR-12-2017-0479
- Pappas, I., Sharma, K., Mikalef, P. and Giannakos, M. (2020), "How quickly can we predict users' ratings on aesthetic evaluations of websites? Employing machine learning on eye-tracking data", paper presented at the Conference on e-Business, e-Services and e-Society, Springer, 429–440. doi: 10.1007/978-3-030-45002-1_37
- Paré, G., Trudel, M-C., Jaana, M. and Kitsiou, S. (2015), "Synthesizing information systems knowledge: A typology of literature reviews", *Information & Management*, Vol. 52 No. 2, pp.183–199. doi: 10.1016/j.im.2014.08.008
- Park, S., Choi, D. and Kim, J. (2004), "Critical factors for the aesthetic fidelity of web pages: Empirical studies with professional web designers and users", *Interacting with Computers*, Vol. 16 No. 2, pp.351-376. doi: 10.1016/j.intcom.2003.07.001
- Pentus, K., Mehine, T. and Kuusik, A. (2014), "Considering emotions in product package design through combining conjoint analysis with psycho physiological measurements", *Procedia - Social and Behavioral Sciences*, Vol 148 No. 10, pp.280–290.
- Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y. Y. and Podsakoff, N. P. (2003), "Common method biases in behavioral research: A critical review of the literature and

- recommended remedies. *Journal of Applied Psychology*, Vol. 88 No. 5, pp.879–903. doi: 10.1037/0021-9010.88.5.879
- Purchase, H., Hamer, J., Jamieson, A. and Ryan, O. (2011), “Investigating objective measures of web page aesthetics and usability”, *Conferences in Research and Practice in Information Technology Series*, Vol. 117, pp.19-28.
- Reinecke, K. and Gajos, K.Z. (2014), “Quantifying visual preferences around the world”, paper presented at the SIGCHI Conference on Human Factors in Computing Systems (CHI '14), ACM, USA, pp.11–20.
- Reinecke, K., Yeh, T., Miratrix, L., Mardiko, R., Zhao, Y., Liu, J. and Gajos, K.Z. (2013), “Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness”, paper presented at the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, pp.2049–2058. doi: 10.1145/2470654.2481281
- Riegler, A. and Holzmann, C. (2018), “Measuring visual user interface complexity of mobile applications with metrics”, *Interacting with Computers*, Vol. 30 No. 3, pp.207–223. doi: 10.1093/iwc/iwy008
- Rogers Y. and Osborne, D.J. (1987), “Pictorial communication of abstract verbs in relation to human-computer interaction”, *British Journal of Psychology*, Vol. 78 No. 1, pp.99–112. doi: 10.1111/j.2044-8295.1987.tb02229.x
- Rui, Z. and Gu, Z. (2021), “A review of EEG and fMRI measuring aesthetic processing in visual user experience research”, *Computational Intelligence and Neuroscience*, Vol 2021. doi: 10.1155/2021/2070209
- Russell, D.W. (2002), “In search of underlying dimensions: The use (and abuse) of factor analysis in personality and social psychology bulletin”, *Personality and Social Psychology Bulletin*, Vol. 28 No. 12, pp.1629–1646. doi: 10.1177/014616702237645
- Salimun, C., Purchase, H.C., Simmons, D. and Brewster, S. (2010), “The effect of aesthetically pleasing composition on visual search performance”, paper presented at the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries, ACM, Reykjavik, Iceland, pp.422-431. doi: 10.1145/1868914.1868963
- Salman, Y.B., Cheng, H.I. and Patterson, P.E. (2012), “Icon and user interface design for emergency medical information systems: A case study”, *International Journal of Medical Informatics*, Vol. 81 No. 1, pp.29–35. doi: 10.1016/j.ijmedinf.2011.08.005
- Salman, Y. B., Kim, Y. and Cheng, H. (2010), “Senior-friendly icon design for the mobile phone”, paper presented at the 6th International Conference on Digital Content, Multimedia Technology and its Applications (IDC 2010), IEEE, Seoul, South Korea, pp.103–108.
- Sarsam, S.M. and Al-Samarräie, H. (2018), “Towards incorporating personality into the design of an interface: A method for facilitating users' interaction with the display”, *User Modeling and User-Adapted Interaction*, Vol. 28 No. 1, pp.75–96. doi: 10.1007/s11257-018-9201-1

- Schifferstein, H. N., Fenko, A., Desmet, P. M., Labbe, D. and Martin, N. (2013), “Influence of package design on the dynamics of multisensory and emotional food experience”, *Food Quality and Preference*, Vol. 27 No. 1, pp.18–25. doi: 10.1016/j.foodqual.2012.06.003
- Selnes, F. and Grønhaug, K. (1986) “Subjective and objective measures of product knowledge contrasted”, *Advances in Consumer Research*, Vol. 13 No. 1, pp.67–71.
- Shaikh, A.D. (2009), “Know your typefaces! Semantic differential presentation of 40 onscreen typefaces”, *Usability News*, Vol. 11, pp.23–65.
- Shu, W. and Lin, C.-S. (2014), “Icon design and game app adoption”, paper presented at the 20th Americas Conference on Information Systems, Georgia, USA.
- Silayoi, P. and Speece, M. (2004), “Packaging and purchase decisions: An exploratory study on the impact of involvement level and time pressure”, *British Food Journal*, Vol. 106 No. 8, pp.607–628. doi: 10.1108/00070700410553602
- Smith, G.E. (1995), “Framing product design: Using design communication to facilitate user learning”, *Journal of Business & Industrial Marketing*, Vol. 10 No. 5, pp.6–21. doi: 10.1108/08858629510103860
- Statista (2016), “Mobile phone gaming penetration in the United States from 2011 to 2020”, <https://www.statista.com/statistics/234649/percentage-of-us-population-that-play-mobile-games/> (accessed January 2, 2022).
- Statista (2021), “Most popular Apple App Store categories in June 2021, by share of available apps”, <https://www.statista.com/statistics/270291/popular-categories-in-the-app-store/> (accessed January 2, 2022).
- Sutcliffe, A. (2002), “Assessing the reliability of heuristic evaluation for web site attractiveness and usability”, paper presented at the 35th Annual Hawaii International Conference on System Sciences, Hawaii, USA, pp. 1838–1847. doi: 10.1109/HICSS.2002.994098
- Tabachnick, B.G. and Fidell, L.S. (2007), *Using multivariate statistics*, Allyn and Bacon/Pearson, Boston.
- Thüring, M. and Mahlke, S. (2007), “Usability, aesthetics and emotions in human–technology interaction”, *International Journal of Psychology*, Vol. 42 No. 4, pp.253–264. doi: 10.1080/00207590701396674
- Timely (2021), “The attention economy: what it is, what it’s doing to you”, available at: <https://memory.ai/timely-blog/the-attention-economy> (accessed January 3, 2022).
- Tractinsky, N. (1997), Aesthetics and apparent usability: Empirically assessing cultural and methodological issues, paper presented at the ACM SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, pp.115–122. doi: 10.1145/258549.258626
- Tractinsky, N., Katz, A.S. and Ikar, D. (2000), “What is beautiful is usable”, *Interacting with Computers*, Vol. 13 No. 2, pp.127–145. doi: 10.1016/S0953-5438(00)00031-X

- Trochim, W.M.K. (2000), *Introduction to validity. Social research methods 2nd ed.*, Cincinnati, Ohio: Atomic Dog Publishing.
- Tuch, A.N., Bargas-Avila, J.A. and Opwis, K. (2010), “Symmetry and aesthetics in website design: It's a man's business”, *Computers in Human Behavior*, Vol. 26 No. 6, pp.1831-1837. doi: 10.1016/j.chb.2010.07.016
- Underwood, R.L., Klein, N.M. and Burke, R.R. (2001), “Packaging communication: attentional effects of product imagery”, *Journal of Product & Brand Management*, Vol. 10 No. 7, pp.403–422. doi: 10.1108/10610420110410531
- Uribe, S., Alvarez, F. and Menéndez, J.M. (2017), “User's web page aesthetics opinion: a matter of low-level image descriptors based on MPEG-7”, *ACM Transactions on the Web*, Vol. 11 No. 1, pp.1–25. doi: 10.1145/3019595.
- Vanderdonckt, J. and Gillo, X. (1994), “Visual techniques for traditional and multimedia layouts”, paper presented at the Workshop on Advanced Visual Interfaces AVI, Bari, Italy, pp.95–104. doi: 10.1145/192309.192334
- van Rompay, T.J.L., Pruyn, A.T.H. and Tieke, P. (2009), “Symbolic meaning integration in design and its influence on product and brand evaluation”, *International Journal of Design*, Vol. 3 No. 2, pp.19–26.
- Voigt, P. and Von dem Bussche, A. (2017). *The EU general data protection regulation (GDPR). A Practical Guide*. Cham: Springer International Publishing
- Wallendorf, M. and Arnould, E.J. (1988), “‘My favorite things’: A cross-cultural inquiry into object attachment, possessiveness, and social linkage”, *Journal of Consumer Research*, Vol. 14 No. 4, pp.531–547.
- Wang, M. and Li, X. (2017), “Effects of the aesthetic design of icons on app downloads: evidence from an android market”, *Electronic Commerce Research*, Vol. 17 No. 1, pp.83–102. doi: 10.1007/s10660-016-9245-4
- Wang, C., Sarcar, S., Kurosu, M., Bardzell, J., Oulasvirta, A., Miniukovich, A. and Ren, X. (2018), Approaching aesthetics on user interface and interaction design, paper presented at the 2018 ACM International Conference on Interactive Surfaces and Spaces, ACM, Tokyo, pp. 481–484. doi: 10.1145/3279778.3279809
- Webster, J. and Watson, R. T. (2002), “Analyzing the past to prepare for the future: Writing a literature review”, *MIS Quarterly*, Vol. 26 No. 2, pp.xiii–xxiii.
- Wiedenbeck, S. (1999), “The use of icons and labels in an end user application program: An empirical study of learning and retention”, *Behavior & Information Technology*, Vol. 18 No. 2, pp.68–82. doi: 10.1080/014492999119129
- Wu, O., Chen, Y., Li, B. and Hu, W. (2011), “Evaluating the visual quality of web pages using a computational aesthetic approach”, paper presented at the 4th International Conference on Web Search and Data Mining, ACM, pp.337–346. doi: 10.1145/1935826.1935883
- Wu, O., Zuo, H., Hu, W. and Li, B. (2016), “Multimodal web aesthetics assessment based on structural svm and multitask fusion learning”, *IEEE Transactions on Multimedia*, Vol. 18 No. 6, pp.1062–1076. doi: 10.1109/TMM.2016.2538722

- Xing B., Si H., Chen J., Ye M. and Shi L. (2021), “Computational model for predicting user aesthetic preference for GUI using DCNNs”, *CCF Transactions on Pervasive Computing and Interaction*, Vol. 3, pp.147–169.
- Yang, Y. and Klemmer, S.R. (2009), “Aesthetics matter: leveraging design heuristics to synthesize visually satisfying handheld interfaces”, extended abstract presented at the SIGCHI Conference on Human Factors in Computing Systems (CHI EA '09), ACM, USA, pp.4183–4188. doi: <https://doi.org/10.1145/1520340.1520637>
- Yun, M.H., Han, S.H., Hong, S.W. and Kim, J. (2003), “Incorporating user satisfaction into the look-and-feel of mobile phone design”, *Ergonomics*, Vol. 46, pp.1423–1440. doi: [10.1080/00140130310001610919](https://doi.org/10.1080/00140130310001610919)
- Zen, M. and Vanderdonckt, J. (2014), “Towards an evaluation of graphical user interfaces aesthetics based on metrics”, paper presented at the IEEE 8th International Conference on Research Challenges in Information Science (RCIS), Marrakech, Morocco, pp.1–6. doi: [10.1109/RCIS.2014.6861050](https://doi.org/10.1109/RCIS.2014.6861050)
- Zen, M. and Vanderdonckt, J. (2016), “Assessing user interface aesthetics based on the inter-subjectivity of judgment”, paper presented at the 30th International BCS Human Computer Interaction Conference. BCS, Swindon, United Kingdom, pp.1–12. doi: [10.14236/ewic/HCI2016.25](https://doi.org/10.14236/ewic/HCI2016.25)

PUBLICATION I

**Evaluation instruments for visual aesthetics of graphical user interfaces:
A review**

Henrietta Jylhä & Juho Hamari (2022)

IEEE Transactions on Visualization and Computer Graphics (in-review)

Publication reprinted with the permission of the copyright holders.

PUBLICATION II

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

Henrietta Jylhä & Juho Hamari (2020)

User Modeling and User-Adapted Interaction, Vol. 30 No. 5, pp. 949–982
DOI: 10.1007/s11257-020-09263-7

Publication reprinted with the permission of the copyright holders.



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 / Published online: 17 May 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.lininfo.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; Vanderdonck and Gillo 1994; Zen and Vanderdonck 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 (Vanderdonck 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 Vanderdonckt 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 Vanderdonckt 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 Vanderdonckt 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 (Vanderdonck 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; Vanderdonck and Gillo 1994), semi-automated computation of user interfaces with the online tool QUESTIM (Zen and Vanderdonck 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
	Higher than master's degree	9	1.6
Employment	Working full-time	133	23.4
	Working part-time	62	10.9
	Student	351	61.7
	Unemployed	11	1.9
	Retired	1	.2
Gender	Male	297	52.2
	Female	257	45.2
	Other	15	2.6
Yearly income	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), Goonetilleke 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), Goonetilleke 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				
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

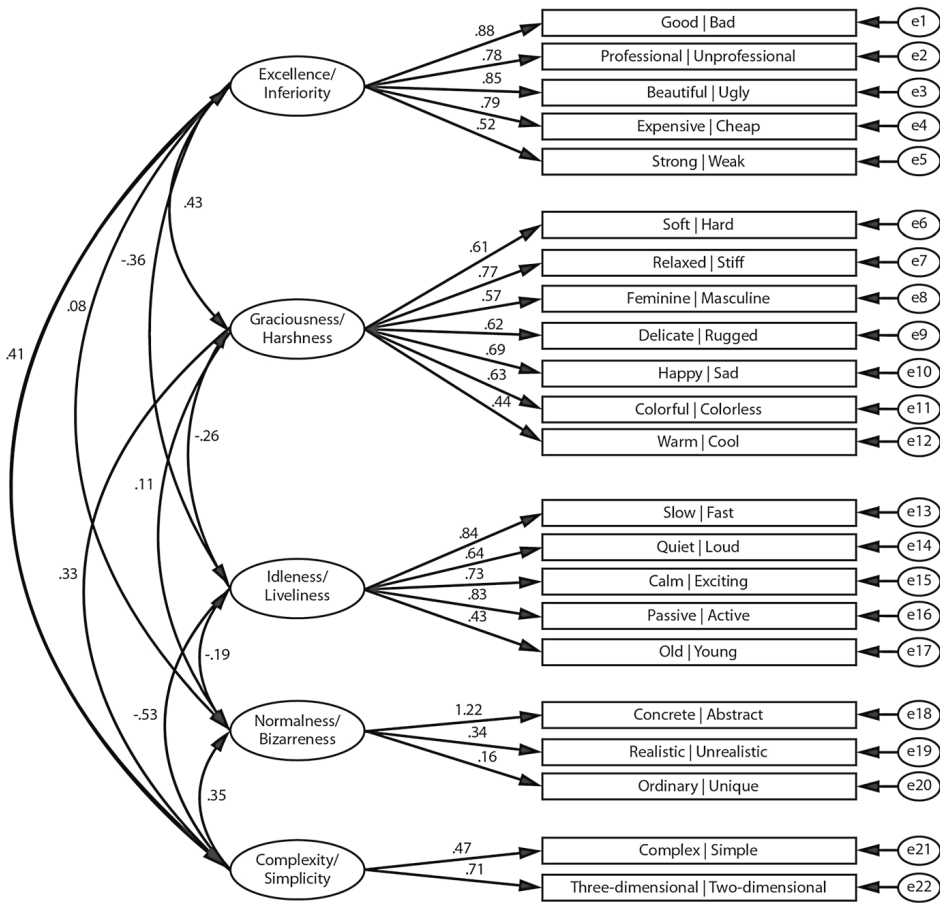


Fig. 1 Initial model with 22 items (standardized weights)

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.

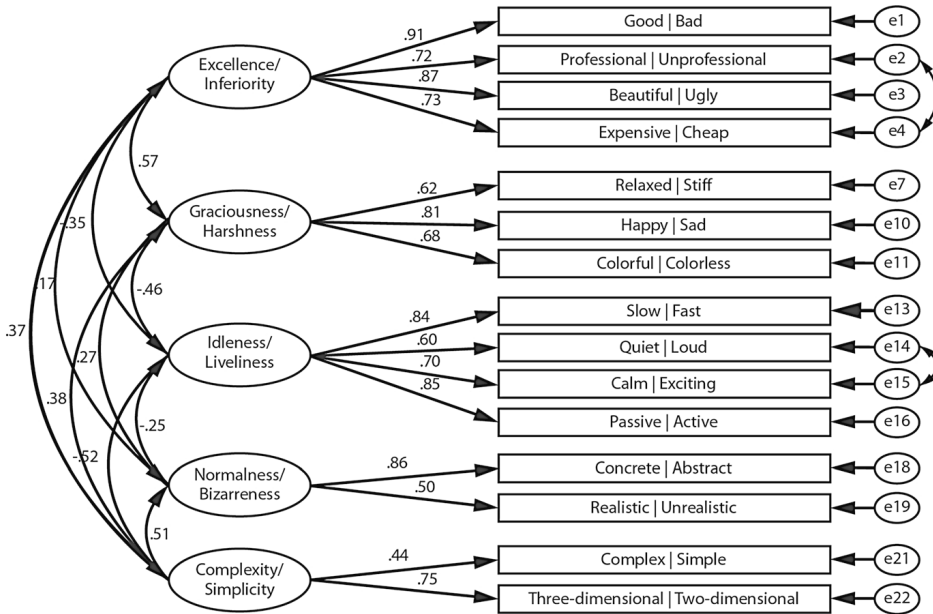


Fig. 2 Adjusted model with 15 items and covaried errors (standardized weights)

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 VISQUAL 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/10.1177/2050157916639348>
- 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., Veyerzer, 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>
- Debevc, 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.ijhcs.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)
- Kurosuo, 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 Bruijin, 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 Bruijin, 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 Bruijin, 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/10.1080/014492999119129>
- 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

PUBLICATION
III

An icon that everyone wants to click: How perceived aesthetic qualities predict app icon successfulness

Henrietta Jylhä & Juho Hamari (2019)

International Journal of Human-Computer Studies, Vol. 130, pp. 73–85

DOI: 10.1016/j.ijhcs.2019.04.004

Publication reprinted with the permission of the copyright holders.



An icon that everyone wants to click: How perceived aesthetic qualities predict app icon successfulness[☆]

Henrietta Jylhä^{a,*}, Juho Hamari^{a,b}

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

^b Gamification Group, Faculty of Humanities, University of Turku, Finland



ARTICLE INFO

Keywords:

Iconography
Aesthetics
Semantic differential
Mobile apps
Graphical user interfaces
Digital marketing

ABSTRACT

Mobile app markets have been touted as fastest growing marketplaces in the world. Every day thousands of apps are published to join millions of others on app stores. The competition for top grossing apps and market visibility is fierce. The way an app is visually represented can greatly contribute to the amount of attention an icon receives and to its consequent commercial performance. Therefore, the *icon* of the app is of crucial importance as it is the first point of contact with the potential user/customer amidst the flood of information. Those apps that fail to arouse attention through their icons danger their commercial performance in the market where consumers browse past hundreds of icons daily. Using semantic differential scale (22 adjective pairs), we investigate the relationship between consumer perceptions of app icons and icon successfulness, measured by 1) overall evaluation of the icon, 2) willingness to click the icon, 3) willingness to download the imagined app and, 4) willingness to purchase the app. The study design was a vignette study with random participant ($n = 569$) assignment to evaluate 4 icons ($n = 2276$) from a total of pre-selected 68 game app icons across 4 categories (concrete, abstract, character and text). Results show that consumers are more likely to interact with app icons that are aesthetically pleasing and convey good quality. Particularly, app icons that are perceived unique, realistic and stimulating lead to more clicks, downloads and purchases.

1. Introduction

After app stores became prominent venues for providing software, the number of mobile apps has been constantly growing at a fast pace (Moreira et al., 2014). Online storefronts try to attract critical masses in various ways, but effective design is necessary for consumer engagement (Overby and Sabyasachi, 2014). Rapid changes in the app markets and consumer mindsets poise new possibilities and challenges in the world-wide competition of commercial success, which motivates the need for further research on app icons and consumer behavior.

App stores house a massive number of mobile applications, also known as apps. To this date, the total number of app downloads from app stores worldwide is estimated 197 billion.¹ Furthermore, global

apps industry revenue has been predicted to rise to 188.9 billion U.S. dollars in 2020.² In light of these statistics, the impact of the apps industry to economic growth is undeniably high. All apps are listed on app stores as icons – a graphic that “provides a quick, intuitive representation of an action, a status or an app”.³ An icon-based graphical user interface (GUI) common to smartphones and tablets has a limited display area, which is why app icons should provide good recognition and user preference (Böhmer and Krüger, 2013; Chen, 2015; Hou and Ho, 2013). Icons essentially act as a first-pass filter for saturated app markets, which is why they need to immediately capture a consumer’s attention.⁴ App icon is in many cases the first and most powerful opportunity to succeed in user engagement on the highly competitive app store markets (Woolridge and Schneider, 2011), hence developers and

[☆] No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work.

* Corresponding author.

E-mail addresses: henrietta.jylha@tuni.fi (H. Jylhä), juho.hamari@tuni.fi (J. Hamari).

URL: <http://gamification.group> (H. Jylhä).

¹ Statista, “Number of mobile app downloads worldwide in 2016, 2017 and 2021 (in billions),” <http://www.statista.com/statistics/234649/percentage-of-us-population-that-play-mobile-games/> (Accessed January 30, 2018).

² Statista, “Worldwide mobile app revenues in 2015, 2016 and 2020 (in billion U.S. dollars),” <https://www.statista.com/statistics/269025/worldwide-mobile-app-revenue-forecast/> (Accessed January 30, 2018).

³ Android Developers, “Iconography,” <http://www.androiddocs.com/design/style/iconography.html> (Accessed January 30, 2018).

⁴ Apple Developers, “App icon,” <https://developer.apple.com/ios/human-interface-guidelines/icons-and-images/app-icon/> (Accessed January 30, 2018).

designers must make a strong impact to prompt consumers to choose to interact with their app instead of the many others.

This observation leads us to the following key research questions: *How do consumer perceptions of app icon aesthetics affect icon successfulness*, namely, what are the aesthetic qualities that are likely to engage consumers into interacting with app icons? Moreover, does app icon appearance affect downloading and purchasing behavior of consumers? This topic is significant for research because minimal attention has been provided to how the visual attributes of apps represented on app stores affect consumer behavior (Wang and Li, 2017; Lin and Chen, 2018). Although icons appear commonly on various interfaces, research examining the determinants of icon appeal is scarce (McDougall et al., 1998, 2016). To our knowledge, no theoretical accounts have been proposed to explain the effects of consumer perceptions on app icon successfulness at the time of this research. Therefore, this study intends to lay the groundwork with potentially far-reaching practical and theoretical implications.

Using semantic differential scale (22 adjective pairs), this exploratory (i.e. non-confirmatory) study investigates the relationship between consumer perceptions of app icons and icon successfulness, measured by 1) overall evaluation of the icon, 2) willingness to click the icon, 3) willingness to download the imagined app and, 4) willingness to purchase the app. The study design is a vignette study, in which participants ($n = 569$) were assigned to evaluate 4 randomized icons from a total of pre-selected 68 game app icons across 4 categories (concrete, abstract, character and text). Game app icons were selected to maximize internal validity. This resulted in a total of 2276 individual icon evaluations. The findings show that consumers are more likely to interact with app icons that are aesthetically pleasing and convey good quality. This contrasts prior research (Shaikh, 2009) on onscreen typeface design and usage, which was implemented as a basis for our experiment. Furthermore, gaps in prior icon design theories are exposed regarding predictors of icon successfulness in terms the adjective pairs Concrete–Abstract and Complex–Simple, as no consistent statistically significant effect was found among them in our study.

2. Background

2.1. Graphical presentation in human–computer interaction

Icons are pictographic symbols of data or processes within a computer system, applied principally to graphics-based interfaces of operating systems (Gittins, 1986). Icons are widely used in human–computer interaction and they have replaced commands and menus as the means by which the computer supports a dialogue with the end-user (García et al., 1994; Gittins 1986; McDougall et al., 1998; Huang et al., 2002). Similar to mobile platforms, iconic interfaces have made their way into our everyday life. Advances in technology result in additional features and further, additional icons. The evolution of icons is traced back to signs (Goonetilleke et al., 2001). Signs are elements that “stand to someone for something in some respect or capacity” (Peirce, 1932). This can be interpreted in the sense that signs as well as icons have a symbolic meaning or connotation behind them. Prior research (Wiedenbeck, 1999) supports this by noting that icons are interface objects that represent a larger system in a simplified, pictorial manner. As we communicate through symbols, these symbols must also be embedded in icons to evoke the desired connotation in the viewer (Horton, 1996).

The terms icon and symbol are differentiated in that icons have a physical connection to a target or function, whereas symbols have an arbitrary, indirect relationship to that which they refer (Horton, 1994). However, the use of the term “icon” to describe symbols has become dominant especially in the interactive field (Horton, 1996). Thus, the everyday usage of “icon” stands for any graphic on an interactive button, and these icons can represent system objects such as files or folders, or actions such as messaging or calling (Wiedenbeck, 1999).

Furthermore, leisurely icons, such as game and movie icons, often depict characters and other relevant features to the title which they represent. This is believed to enhance product identity and brand personality (Phillips, 1996).

The reason why icons are extensively used is due to many factors. Icons facilitate human–computer interaction because they are swiftly recognized and memorized (Horton, 1994, 1996; McDougall et al., 1999; Wiedenbeck, 1999). Icons are also more convenient for universal communication than text, since language interpretation is not an obstacle (Arend et al., 1987; Horton, 1994, 1996; Lodding, 1983; McDougall et al., 1999). Despite these positive results of icon usage, there is little published research on app icons, justifying further investigation.

One aspect of prior research on icon aesthetics concerns whether concrete or abstract icons are more effective from user perspective (e.g. 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., 1998, 1999, 2000; Moyes and Jordan, 1993; Rogers and Osborne, 1987). Icon concreteness is the extent to which it depicts real objects (Isherwood et al., 2007), whereas icon abstractness tends to have less obvious connections with real objects (McDougall et al., 1999). Some studies (e.g. Hou and Ho, 2013; McDougall and Reppa, 2008; McDougall et al., 1999, 2000; Moyes and Jordan, 1993; Rogers and Osborne, 1987) show that most users prefer concrete, rich icon designs to abstract, simplified icons, while others have found conflicting results (Arend et al., 1987; McDougall et al., 1998). Prior research has also suggested that concreteness may not be of primary importance after all, rather semantic distance and familiarity may be more important (Blankenberger and Hahn, 1991; Dewar, 1999; Isherwood et al., 2007; McDougall et al., 1998, 1999; McDougall and Reppa, 2008; Schröder and Ziefle, 2008). Furthermore, icon familiarity has been acknowledged to help reduce the amount of information to communicate a message (Arab et al., 2013; Forsythe et al., 2008) which makes an icon easier to understand.

The juxtaposition of concrete and abstract icons is sometimes referred to as the *guessability gulf* (Moyes and Jordan, 1993). This is because concrete icons are easier to cognitively process at first sight than abstract icons. Despite the debate between concreteness and abstractness of icons, it is noteworthy that icon preference is affected by many factors. Computer icons have evolved from information signs to a part of consumer culture (Huang et al., 2002), therefore different types of icons may be suitable for different purposes and personalities. For example, concrete icons can be useful in public information systems or warnings (McDougall et al., 1998; McDougall and Reppa, 2013) where the goal is to clearly communicate information, whereas more stylistic icon design may promote other ends (Hou and Ho, 2013).

Another aspect of effective icon design is the speed and ease with which icons can be understood (e.g. Arend et al., 1987; Blankenberger and Hahn, 1991; Dewar, 1999; Lodding, 1983; Isherwood et al., 2007; McDougall and Reppa, 2008; McDougall and Reppa, 2013; McDougall et al., 2016; Moyes and Jordan, 1993; Wiedenbeck, 1999). Prior research (McDougall and Reppa, 2013; McDougall et al., 2016) on interface icon design has found that processing fluency affects icon appeal and that simple icon design has been shown to lead to user satisfaction. Factors influencing icon processing are e.g. icon familiarity and complexity, meaning that the easier the icon is to process due to simple design and earlier experience with similar icons, the more appealing it is (Arab et al., 2013; Choi and Lee, 2012; Dewar, 1999; Forsythe et al., 2008; Goonetilleke et al., 2001; Huang et al., 2002; Isherwood et al., 2007; McDougall and Reppa, 2008; McDougall et al., 1998, 2016; McDougall and Reppa, 2013; Moyes and Jordan, 1993).

Further concerning icon design and icon effectiveness, it has been speculated by prior research that the most important features of an icon are dominance, uniqueness or discriminability, and unambiguity (Arend et al., 1987; Dewar, 1999; Goonetilleke et al., 2001; Huang et al., 2002). Dominance is defined as a characteristic intrinsic to a

function and its context. An icon element is dominant if the other icon elements can be inferred from the first one. Therefore, icons with dominant elements are said to be processed more fluently than icons with redundant elements (Goonetilleke et al., 2001). An icon is said to convey uniqueness or discriminability when the representation and its function has perceptual immediacy, making an icon distinguishable and locatable among other icons (Goonetilleke et al., 2001; Huang et al., 2002). For example, icons featuring elements that are not interchangeable with other representations enhance icon uniqueness. The use of unique, visually distinctive icons has been shown to lead into better performance compared to icons that are not perceived as unique (Arend et al., 1987; Goonetilleke et al., 2001; Huang et al., 2002). However, uniqueness/discriminability has not been defined formally as specific differences in icons but rather it is in the eye of the beholder. Therefore, there is not a tangible way to define what it means in terms of specified features of an icon. Unambiguity is defined as a representation that can be associated with only one of the functions in a given context (Goonetilleke et al., 2001). If an icon is ambiguous, i.e. it holds multiple meanings in a single context, it may result in various interaction problems to users, especially if they have limited experience in icon identification (Black, 2017; Goonetilleke et al., 2001; Rogers and Osborne, 1987; Salman et al., 2010). For example, an icon depicting only a human face with an open mouth can refer to various actions such as eating, drinking or speaking, and therefore the representation is ambiguous. In this case, adding another element guiding the contextual function would aid in reaching unambiguity. Such an element could be e.g. a cup of coffee, which would make the icon easier to interpret as something related to the function of drinking (Goonetilleke et al., 2001).

App icons are a necessary part of branding and product design, as icons are key marketing elements presented to the consumer before downloading an app (i.e. product). Effective package and product design has been widely acknowledged as a factor for advantage in economic competition (e.g. Ares et al., 2011; Creusen and Schoormans, 2005; Creusen et al., 2010; Rundh, 2009; Schifferstein et al., 2013). Consumers use a lot of time and effort to evaluate how a product is presented, and often form their perceptions on brands based on design (Orth and Malkewitz, 2008). Hence, design affects brand and product selection and may drive purchase decisions (Ares et al., 2011; Creusen and Schoormans, 2005; Creusen et al., 2010; Fenko et al., 2010; Orth and Malkewitz, 2009; Schifferstein et al., 2013; van Rompay et al., 2009). In this light, the effects of product design should be of great importance to app designers, marketers and developers.

Product presentation can be divided into two main categories, visual and informational elements. Visual elements (i.e. graphics) include layout, color, typography, size and shape, whereas informational elements include written information about the product (Silayoi and Speece, 2004). App icons belong to the category of visual elements that communicate to the consumer most directly. In decision-making, consumers spontaneously form impressions of product content quality based on how a product is presented (Underwood et al., 2001; Yun et al., 2003) and these impressions can have lasting impact. For example, consumers perceive highly saturated colors as exciting (Labrecque and Milne, 2011), making them popular in product presentation. Furthermore, it is believed that effective visual elements in product presentation evoke more of an emotional response than informational elements (Silayoi and Speece, 2004), which in turn brings extra value to the product and increases the possibility of purchase (Cho and Lee, 2005). Hence, emotional impact is important when creating products, services and brands (Crossley, 2003).

2.2. Related work

Prior studies have investigated effective icon design in terms of icon concreteness and abstractness (e.g. 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), icon familiarity (Arab et al., 2013; Dewar, 1999; Forsythe et al., 2008; Huang et al., 2002; Isherwood et al., 2007; McDougall and Reppa 2008, McDougall and Reppa, 2013; McDougall et al., 2016; Moyes and Jordan, 1993), icon simplicity and complexity (Choi and Lee, 2012; Goonetilleke et al., 2001; McDougall and Reppa, 2008, McDougall and Reppa, 2013; McDougall et al., 2016) as well as uniqueness/discriminability and unambiguity (Creusen and Schoormans, 2005; Creusen et al., 2010; Dewar, 1999; Goonetilleke et al., 2001; Huang et al., 2002; Salman et al., 2010). However, app icons on app stores differ from other interface icons in that they are not only designed for interaction, but also as marketing assets meant to attract the consumer and to stand out in a display of many other offerings, much like any other, more tangible product. From this perspective, there are a plethora of other aspects that could also prove to be important determinant of app icon successfulness.

The handful of studies that have investigated the relationship between app icons and consumer behavior have consensus in that app icons play an important role in the mobile app markets, and that attractive icons have the power to trigger the interest of consumers (Burgers et al., 2016; Chen, 2015; Hou and Ho, 2013; Lin and Chen, 2018; Salman et al., 2010; Shu and Lin, 2014; Wang and Li, 2017). Nevertheless, mixed results have been reported on the attributes of successful icons. Studies that have investigated app icon design to understand task performance and user preference of different icon types have found that consumers prefer detailed, pictorial app icon design (Chen, 2015; Hou and Ho, 2013), sometimes regardless of inefficiency on task performance (Chen, 2015). Other findings on the visual attributes of app icon appearance recommend simplicity and complexity to be balanced for consumer appeal, as well as adding slight asymmetry to the design (Wang and Li, 2017). Moreover, positive evidence suggests that color is an important aspect of app icon design, as particularly bright and colorful icons increase the chance of app downloads and consumer preference (Salman et al., 2010; Wang and Li, 2017). Prior research on the relationship between icon attributes and consumer choice of apps has reported that app icon successfulness is dependent on app type as well as user personality and demographics (Hou and Ho, 2013; Salman et al., 2010; Shu and Lin, 2014), which complicates conclusions on the topic. Conflicting findings may be due to the fact that aesthetic appeal is a multi-dimensional topic that consists of various dimensions (Reppa and McDougall, 2015). According to the literature review herein, it appears that currently there does not exist a coherent body of knowledge on the issue of understanding how icon aesthetics affect perception and behavior. This is especially so as there exist only few studies on the topic as well as because their results are slightly mixed and conflicting. No clear trajectory of results emerges from the literature. Apps are used for several purposes by users with different profiles, thus it is important to advance knowledge in this topic to avoid pitfalls during icon design. As literature on this topic is limited, further investigation is justified.

Therefore, we set out to explore the relationship between consumer perceptions of app icon aesthetics and icon successfulness, to find out what are the perceived aesthetic qualities that are likely to engage consumers into interacting with app icons to fill the gap in prior literature identified above. Furthermore, we wish to further the body of research so that a pathway to a more coherent conclusion of icon aesthetics and consumer perceptions could be formed. The following section introduces the study design.

3. Methods and data

In order to find out how consumer perceptions of app icon aesthetics affect icon successfulness, we employed a semantic differential scale of 22 adjective pairs, measured by 1) overall evaluation of the icon, 2) willingness to click the icon, 3) willingness to download the imagined

Table 1
Demographic information.

		n	%
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	0.9
	High school	135	23.7
	College	95	16.7
	Bachelor's degree	227	39.9
	Master's degree	98	17.2
	Higher than master's degree	9	1.6
Employment	Working full-time	133	23.4
	Working part-time	62	10.9
	Student	351	61.7
	Unemployed	11	1.9
	Retired	1	0.2
Game apps downloaded (per week)	0	429	75.4
	1	104	18.3
	2	14	2.5
	3	9	1.6
	4	2	0.4
	5	1	0.2
	Missing	10	1.8
Gender	Male	297	52.2
	Female	257	45.2
	Other	15	2.6
Yearly income	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

app and, 4) willingness to purchase the app. This study utilized a within-subjects vignette approach, where each subject ($n = 569$) served in four treatments. Participants were assigned to evaluate 4 randomized icons from a total of pre-selected 68 game app icons across 4 categories (concrete, abstract, character and text) in a hypothetical situation setting instead of a description more typical to vignette studies. The aim was to acquire reliable data by exposing the participants close to a realistic setting outside the app store context. This resulted in a total of 2276 individual icon evaluations. The experiment was carried out as a self-administered online task. The following Section 3.1 describes the participants in the study.

3.1. Participants

The sample is composed of a nonprobability convenience sample with 569 respondents who participated in the study and assessed game app icons through the vignette experiment. A link to the online experiment was advertised in Facebook groups and Finnish student organizations' mailing lists. The participants predominantly resided in Finland (92.8%). Other countries clearly represented in the data were the United States (2.1%) and United Kingdom (2.1%). Please refer to Table 1 for demographic details of participants.

The gender split across participants was rather equal, as only slightly more than half were male (52.2%). The mean age was 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%). On a weekly basis, most participants (75.4%) did not download any game apps. Missing data (1.8%) was encountered for this item, as the frequency of mobile game downloads was only asked of those who use a smartphone. To counter possible bias in the experiment, participants who did not download game apps frequently were instructed to answer based on their expectations of game app icons they might interact with. Two participants were randomly chosen and awarded a prize (Polar Loop 2 Activity Tracker). No other participation fees were paid. Participants were informed of the purpose of the study and assured anonymity.

3.2. Materials

Sixty-eight game app icons from Google Play Store were selected for the study. The decision to narrow down the sample to game app icons was made to eliminate further variability that might stem from the nature of the app and thus increase internal validity of the experiment, but also external validity in terms of results applied to the game icons. In order to avoid any systematic bias, 4 icons corresponding to dominant icon styles (concrete, abstract, character and text) were selected from each of 17 categories for game apps (action, adventure, arcade, board, card, casino, casual, educational, music, puzzle, racing, role playing, simulation, sports, strategy, trivia and word). Because icon design for app stores is category-dependent (Shu and Lin, 2014), we considered it justified to include icons from all categories. Prior literature highlights the relevance of concreteness and abstractness in icon design (e.g. 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), hence they were included in this experiment. The icons are presented in Table 2.

Looking at the icons on app stores, characters and typography are prevalent elements usually seen on app icons. It has been argued that faces on app icons are widely used because of the immediate impact and memorability they have due to neural processing of facial expressions.⁵ Furthermore, as the study design is based on prior research (Shaikh, 2009) on onscreen typeface and usage, text elements were included. During the selection phase we ensured that one icon from each category was dominantly characteristic of one of these 4 attributes. Since the categories are overlapping to an extent, separation between the categories was based on the most prominent elements in the icons. For example, icons in the "concrete" genre were selected in such a way that facial structures were not dominant in the icon, whereas in the "character" genre, the main element in the selected icons was a close-up image where the facial expression was prevalent.

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 3rd to 17th 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 as visually rich sample of icons as possible, meaning that several different computer graphic techniques were included, such as 2D and 3D rendered images.

3.3. Measurements

Semantic differential scale was used to measure respondent

⁵ Chartboost, "Power-Up Report – July 2015," <https://chartboost.s3.amazonaws.com/blog/power-up-report-july-2015-building-an-empire-mobile-strategy-games.pdf> (Accessed September 14, 2018).

Table 2
Icons in the study.

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

Table 3
Adjectives, means and standard deviations.

Adjective pairs	Mean	Std.
Beautiful–Ugly	4.57	1.618
Expensive–Cheap	4.83	1.563
Good–Bad	4.34	1.641
Happy–Sad	3.80	1.507
Hard–Soft	3.81	1.545
Strong–Weak	3.93	1.464
Feminine–Masculine	4.34	1.388
Delicate–Rugged	4.42	1.368
Relaxed–Stiff	4.47	1.560
Old–Young	3.98	1.611
Passive–Active	3.97	1.708
Slow–Fast	3.87	1.576
Calm–Exciting	3.96	1.452
Cool–Warm	3.97	1.436
Quiet–Loud	4.12	1.601
Adjective pairs related to aesthetic qualities		
Concrete–Abstract	4.03	1.998
Professional–Unprofessional	4.22	1.736
Unique–Ordinary	4.60	1.651
Colorful–Colorless	3.77	1.810
Realistic–Unrealistic	4.22	1.592
Two-dimensional–Three-dimensional	3.33	1.863
Complex–Simple	4.69	1.669

evaluations of aesthetic aspects of the icons. A total of 22 adjective pairs was formulated and assigned to each icon. The polarity of the adjective pairs was reversed so that perceivably positive and negative adjectives did not align on the same side of the scale. All of the adjective pairs were chosen according to prior research (Shaikh, 2009) on onscreen typeface design and usage. Additionally, adjectives related to icons were added as suggested per previous literature on effective icon design. 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, 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 were added to specifically measure the aesthetics of the icons include professional and unprofessional, colorful and colorless, realistic and unrealistic as well as two-dimensional and three-dimensional.

Table 3 lists the adjective pairs used in the study and presents an overview of the means and standard deviations. There were no outlier values and the range between the lowest and highest scores cluster closer to the average even though the 68 icons were quite different from each other. All the mean scores were between 3.5 and 4.5 for each evaluation. This indicates little skewness in the data.

To measure participants willingness to interact with the icons presented to them, a seven-point Likert scale was utilized to measure the degree of disagree-agreement of the respondents with respect to the likelihood of them clicking, downloading, and purchasing the imagined app behind the icon with an instruction title: “Overall evaluation (judging by the icon alone)” followed by questions: “Compared to the mobile game icons I usually click, I would click this icon”, “Compared to the icons of mobile games I usually download, I would click this icon” and “Compared to the icons of mobile games I usually purchase, I would click this icon.” Respondents were provided the following options on the seven-point scale: “Strongly disagree”, “Disagree”, “Somewhat disagree”, “Neither agree nor disagree”, “Somewhat agree”, “Agree” and “Strongly agree”. Moreover, respondents were asked to give an overall evaluation score for the design of each icon by grading them on a seven-point scale to further assess consumer perceptions of

Table 4
The relationship between consumer perceptions of icons and the willingness to click, download and purchase.

	Beta	<i>p</i>	Beta	<i>p</i>	Beta	<i>p</i>	Beta	<i>p</i>
	Evaluation (<i>R</i> ² = 0.658)		Click (<i>R</i> ² = 0.550)		Download (<i>R</i> ² = 0.530)		Purchase (<i>R</i> ² = 0.425)	
Beautiful–Ugly	-0.246**	0.000	-0.256**	0.000	-0.222**	0.000	-0.201**	0.000
Good–Bad	-0.332**	0.000	-0.357**	0.000	-0.351**	0.000	-0.303**	0.000
Unique–Ordinary	-0.071**	0.000	-0.112**	0.000	-0.098**	0.000	-0.113**	0.000
Hard–Soft	0.049**	0.004	0.055**	0.008	0.056**	0.009	0.054*	0.020
Calm–Exciting	0.072**	0.000	0.069**	0.002	0.086**	0.000	0.049*	0.043
Passive–Active	0.057**	0.004	0.084**	0.000	0.049*	0.048	0.029	0.276
Realistic–Unrealistic	-0.002	0.888	-0.048**	0.007	-0.052**	0.004	-0.060**	0.002
Quiet–Loud	-0.013	0.462	-0.057**	0.007	-0.053*	0.016	-0.051*	0.033
Colorful–Colorless	-0.036*	0.032	0.051*	0.014	0.030	0.156	0.053*	0.021
Feminine–Masculine	0.081**	0.000	0.044*	0.027	0.037	0.068	0.021	0.328
Two–Three-dimensional	0.031*	0.036	-0.050**	0.006	-0.029	0.113	-0.007	0.719
Old–Young	0.043**	0.004	0.020	0.256	0.027	0.147	0.014	0.285
Professional–Unprofessional	-0.126**	0.000	-0.029	0.219	-0.048	0.051	-0.048	0.069
Relaxed–Stiff	-0.055**	0.002	-0.013	0.554	-0.033	0.137	-0.035	0.148
Strong–Weak	-0.060**	0.000	-0.027	0.194	-0.012	0.564	-0.020	0.396
Happy–Sad	0.002	0.907	0.023	0.275	0.042	0.053	0.059*	0.012
Concrete–Abstract	0.024	0.118	0.015	0.413	0.031	0.103	0.039	0.057
Complex–Simple	0.004	0.800	-0.007	0.688	0.008	0.664	0.001	0.954
Cool–Warm	0.000	0.985	0.010	0.569	-0.002	0.911	0.013	0.489
Delicate–Rugged	-0.003	0.832	0.008	0.672	0.011	0.595	-0.001	0.980
Expensive–Cheap	-0.032	0.120	-0.005	0.829	-0.033	0.188	-0.025	0.354
Slow–Fast	-0.018	0.354	0.015	0.547	0.015	0.547	0.043	0.110

p* < 0.05, *p* < 0.01, statistically significant effects bolded.

icon successfulness.

3.4. Procedure

The data was 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 was asked to evaluate game app icons using seven-point 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 an initial “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. Last, 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 was analyzed with IBM SPSS

Statistics version 24 and Microsoft Office Excel 2016. The following section describes the results of the analysis.

4. Results

Multiple linear regression analyses (MLRA) were performed to investigate the relationships between perceptions of icons (represented by the 22 adjective pairs) and each of the four variables related to icon successfulness (1. overall evaluation of the icon, 2. willingness to click the icon, 3. willingness to download the imagined app and, 4. willingness to purchase the imagined app). Please refer to Table 4 for results.

We tested for multicollinearity with variance inflation factors (VIF), a common procedure in regression analysis to observe whether some relationships are masked due to collinearity. Multicollinearity causes inflation in the variances of regression coefficients, which may lead in unreliable conclusions about the relationship between variables. The VIF values for each adjective pair were between 1.5 and 2.7, except for the adjective pair Beautiful–Ugly and Good–Bad (VIF > 3). Please refer to Table 5 for VIF scores.

According to Montgomery and Peck (1992) a VIF value that exceeds 5 (or in some cases 10) implies multicollinearity. In this light, the values in the analysis are more than acceptable. Nevertheless, compared to the other values, the higher VIF of Beautiful–Ugly and Good–Bad may suggest some multicollinearity, albeit that the values are not critically high. However, omitting variables due to relatively high VIF values (in comparison with other variables in the models) is a standard procedure that can be performed as a theory-driven decision. In this study, we aim to make predictions on the more underlying elements of icon aesthetics than those that are conceptually overlapping (i.e. Beautiful–Ugly and Good–Bad) at a higher level. Hence, we considered it worth finding out if there are significant elements hidden in the model when the adjective pairs of the highest VIF scores are removed. Thus, we ran additional post-hoc regression analyses excluding adjective pairs Beautiful–Ugly and Good–Bad. The analyses were performed with the remaining 20 adjective pairs and each of the four variables related to icon successfulness (1. overall evaluation of the icon, 2. willingness to click the icon, 3. willingness to download the imagined app and, 4. willingness to purchase the imagined app). Please refer to Table 6 for results.

Our predictions regarding hidden relationships between variables

Table 5
VIF values.

	VIF
Beautiful–Ugly	3.206
Good–Bad	3.494
Unique–Ordinary	1.326
Hard–Soft	1.924
Calm–Exciting	2.085
Passive–Active	2.570
Realistic–Unrealistic	1.368
Quiet–Loud	2.033
Colorful–Colorless	1.899
Feminine–Masculine	1.730
Two–Three-dimensional	1.443
Old–Young	1.420
Professional–Unprofessional	2.549
Relaxed–Stiff	2.065
Strong–Weak	1.922
Happy–Sad	1.963
Concrete–Abstract	1.503
Complex–Simple	1.338
Cool–Warm	1.350
Delicate–Rugged	1.760
Expensive–Cheap	2.725
Slow–Fast	2.579

due to multicollinearity were supported by the results of the post-hoc analyses (Table 6). Hidden significant effects were found when the analyses were performed without the adjective pairs Beautiful–Ugly and Good–Bad. This is probably caused by the general nature of the adjective pairs that may cause some of the relevant effects to remain undetected when they are kept in the model. Thus, in future research, the initial model could be corrected in such a way that the adjective pairs Beautiful–Ugly and Good–Bad are removed, as they may bias relationships with other variables.

The results indicate that regarding the relationship between consumer perceptions of app icons and their overall evaluation, the following ends of the semantic differentials positively predicted their grade (Table 4): *beautiful, good, unique, soft, exciting, active, colorful, masculine, three-dimensional, young, professional, relaxed and strong*. Without the adjective pairs Beautiful–Ugly and Good–Bad, the following adjectives also revealed to be positive predictors of the overall evaluation (Table 6): *expensive, quiet, realistic, happy, and simple*.

The following ends of the semantic differentials positively predicted

the willingness to click app icons (Table 4): *beautiful, good, unique, soft, exciting, active, realistic, quiet, colorless, masculine and two-dimensional*. Without the adjective pairs Beautiful–Ugly and Good–Bad, the following adjectives also revealed to be positive predictors of the willingness to click (Table 6): *professional, expensive, strong, relaxed, happy, and young*.

The following ends of the semantic differentials positively predicted the willingness to download the imagined app that the icon belongs to (Table 4): *beautiful, good, unique, soft, exciting, active, realistic and quiet*. Without the adjective pairs Beautiful–Ugly and Good–Bad, the following adjectives also revealed to be positive predictors of the willingness to download (Table 6): *professional, expensive, strong, happy, young, and simple*.

The following ends of the semantic differentials positively predicted the willingness to purchase the imagined app that the icon belongs to (Table 4): *beautiful, good, unique, soft, exciting, realistic, quiet, colorless and sad*. Without the adjective pairs Beautiful–Ugly and Good–Bad, the following adjectives also revealed to be positive predictors of the willingness to purchase (Table 6): *professional, expensive, strong, relaxed, and fast*.

Additionally, both of the previous multiple linear regression models (Tables 4 and 6) were performed with stepwise method to acquire a more thorough understanding of the perceptions of icons (represented by the 22 adjective pairs) and each of the four variables related to icon successfulness (1. overall evaluation of the icon, 2. willingness to click the icon, 3. willingness to download the imagined app and, 4. willingness to purchase the imagined app). The purpose of these analyses was to observe whether there are differences in which adjective pairs are inserted in the models, in order to compare to the previous MLRA models. Please refer to Tables 7 and 8 for results.

The stepwise regressions inserted 12 to 9 of the original 22 adjective pairs in the final models per dependent (overall evaluation: 12, willingness to click: 11, willingness to download: 12, and willingness to purchase: 9). Nearly all variables in the models (except the adjective pair Happy–Sad within the willingness to download in Table 8) were statistically significant at the 0.05 or 0.01 level. These findings support the results of the previous regression analyses (Table 4) in that the adjective pairs inserted in the final models of the stepwise analyses were nearly identical to the previous MLRA.

In order to compare the results to the post-hoc MLRA (Table 6), we ran the stepwise regressions without the adjective pairs Beautiful–Ugly and Good–Bad to find out if hidden relationships remained. Please refer

Table 6

The relationship between consumer perceptions of icons and the willingness to click, download and purchase (excluding Beautiful–Ugly and Good–Bad).

	Beta Grade ($R^2 = 0.567$)	<i>p</i>	Beta Click ($R^2 = 0.521$)	<i>p</i>	Beta Download ($R^2 = 0.506$)	<i>p</i>	Beta Purchase ($R^2 = 0.408$)	<i>p</i>
Unique–Ordinary	-0.101**	0.000	-0.143**	0.000	-0.128**	0.000	-0.139**	0.000
Professional–Unprofessional	-0.290**	0.000	-0.204**	0.000	-0.212**	0.000	-0.192**	0.000
Expensive–Cheap	-0.182**	0.000	-0.165**	0.000	-0.182**	0.000	-0.157**	0.000
Hard–Soft	0.065**	0.001	0.071**	0.002	0.071**	0.002	0.067**	0.006
Strong–Weak	-0.148**	0.000	-0.120**	0.000	-0.100**	0.000	-0.097**	0.000
Relaxed–Stiff	-0.111**	0.000	-0.072**	0.002	-0.088**	0.000	-0.083**	0.001
Quiet–Loud	-0.084**	0.000	-0.133**	0.000	-0.123**	0.000	-0.112**	0.000
Calm–Exciting	0.100**	0.000	0.099**	0.000	0.114**	0.000	0.073**	0.004
Realistic–Unrealistic	-0.050**	0.002	-0.099**	0.000	-0.100**	0.000	-0.102**	0.000
Passive–Active	0.078**	0.000	0.106**	0.000	0.070**	0.009	0.048	0.090
Happy–Sad	-0.101**	0.000	-0.086**	0.000	-0.062**	0.006	-0.032	0.185
Old–Young	0.060**	0.000	0.039*	0.048	0.043*	0.028	0.029	0.167
Colorful–Colorless	-0.042*	0.027	0.045*	0.047	0.025	0.281	0.048*	0.046
Two–Three-dimensional	0.038*	0.022	-0.042*	0.032	-0.023	0.255	-0.001	0.951
Complex–Simple	0.039*	0.014	0.031	0.106	0.043*	0.025	0.032	0.113
Feminine–Masculine	0.065**	0.000	0.027	0.214	0.023	0.295	0.008	0.715
Slow–Fast	0.000	0.982	0.034	0.192	0.033	0.212	0.059*	0.036
Concrete–Abstract	0.029	0.084	0.021	0.293	0.036	0.073	0.044	0.042
Cool–Warm	0.006	0.721	0.016	0.395	0.004	0.847	0.018	0.364
Delicate–Rugged	-0.016	0.384	-0.005	0.831	-0.001	0.969	-0.011	0.635

* $p < 0.05$, ** $p < 0.01$, statistically significant effects bolded.

Table 7
Overall evaluation and the willingness to click (stepwise).

Step #	Beta Evaluation ($R^2 = 0.656$)	p	Step #	Beta Click ($R^2 = 0.490$)	p
1 Good–Bad	–0.337**	0.000	1 Good–Bad	–0.376**	0.000
2 Beautiful–Ugly	–0.253**	0.000	2 Beautiful–Ugly	–0.273**	0.000
3 Passive–Active	0.049**	.004	3 Unique–Ordinary	–0.122**	0.000
4 Professional–Unprofessional	–0.142**	0.000	4 Passive–Active	0.095**	0.000
5 Unique–Ordinary	–0.079**	0.000	5 Colorful–Colorless	0.050**	0.008
6 Calm–Exciting	0.066**	0.000	6 Calm–Exciting	0.074**	0.000
7 Old–Young	0.047**	.001	7 Quiet–Loud	–0.053**	0.009
8 Feminine–Masculine	0.071**	0.000	8 Feminine–Masculine	0.052**	0.004
9 Relaxed–Stiff	–0.061**	0.000	9 Two–Three-dimensional	–0.047**	0.006
10 Strong–Weak	–0.060**	0.000	10 Realistic–Unrealistic	–0.040*	0.012
11 Hard–Soft	0.052**	.001	11 Hard–Soft	0.043*	0.015
12 Two–Three-dimensional	0.028*	.036			

* $p < 0.05$.
** $p < 0.01$.

Table 8
The willingness to download and purchase (stepwise).

Step #	Beta Download ($R^2 = 0.466$)	p	Step #	Beta Purchase ($R^2 = 0.372$)	p
1 Good–Bad	–0.363**	0.000	1 Good–Bad	–0.324**	0.000
2 Beautiful–Ugly	–0.233**	0.000	2 Beautiful–Ugly	–0.217**	0.000
3 Unique–Ordinary	–0.104**	0.000	3 Unique–Ordinary	–0.134**	0.000
4 Calm–Exciting	0.088**	0.000	4 Happy–Sad	0.056*	0.011
5 Professional–Unprofessional	–0.058**	0.007	5 Slow–Fast	0.066**	0.000
6 Quiet–Loud	–0.055**	0.008	6 Realistic–Unrealistic	–0.047**	0.007
7 Passive–Active	0.053*	0.011	7 Hard–Soft	0.058**	0.002
8 Happy–Sad	0.038	0.056	8 Professional–Unprofessional	–0.062**	0.008
9 Realistic–Unrealistic	–0.050**	0.004	9 Colorful–Colorless	0.051*	0.016
10 Concrete–Abstract	0.042*	0.021			
11 Hard–Soft	0.053**	0.004			
12 Feminine–Masculine	0.043*	0.021			

* $p < 0.05$.
** $p < 0.01$.

to Tables 9 and 10 for results.

The stepwise regressions that were performed without the adjective pairs Beautiful–Ugly and Good–Bad inserted 15 to 11 adjective pairs in the final models per dependent (overall evaluation: 15, willingness to click: 13, willingness to download: 14, and willingness to purchase: 11). All variables in the models were statistically significant at the 0.05 or 0.01 level. The findings repeat our notion regarding the adjective pairs

Beautiful–Ugly and Good–Bad, namely, that several underlying significant effects are revealed without these two adjective pairs with the highest VIF scores (Table 5) in the model.

The stepwise regression results indicate that regarding the relationship between consumer perceptions of app icons and their overall evaluation, the following ends of the semantic differentials positively predicted their grade (Table 7): *good*, *beautiful*, *active*, *professional*,

Table 9
Overall evaluation and the willingness to click excluding Beautiful–Ugly and Good–Bad (stepwise).

Step #	Beta Evaluation ($R^2 = 0.566$)	p	Step #	Beta Click ($R^2 = 0.387$)	p
1 Professional–Unprofessional	–0.293**	0.000	1 Expensive–Cheap	–0.170**	0.000
2 Happy–Sad	–0.096**	0.000	2 Professional–Unprofessional	–0.203**	0.000
3 Expensive–Cheap	–0.189**	0.000	3 Unique–Ordinary	–0.147**	0.000
4 Passive–Active	0.084**	0.000	4 Happy–Sad	–0.075**	0.000
5 Unique–Ordinary	–0.104**	0.000	5 Passive–Active	0.127**	0.000
6 Relaxed–Stiff	–0.117**	0.000	6 Realistic–Unrealistic	–0.089**	0.000
7 Strong–Weak	–0.148**	0.000	7 Quiet–Loud	–0.128**	0.000
8 Old–Young	0.063**	0.000	8 Strong–Weak	–0.126**	0.000
9 Calm–Exciting	0.099**	0.000	9 Calm–Exciting	0.107**	0.000
10 Quiet–Loud	–0.085**	0.000	10 Relaxed–Stiff	–0.067**	0.002
11 Hard–Soft	0.070**	0.000	11 Complex–Simple	0.040*	0.031
12 Feminine–Masculine	0.059**	0.001	12 Hard–Soft	0.065**	0.003
13 Realistic–Unrealistic	–0.046**	0.002	13 Two–Three-dimensional	–0.047*	0.015
14 Colorful–Colorless	–0.043*	0.021			
15 Complex–Simple	0.032*	0.038			

* $p < 0.05$.
** $p < 0.01$.

Table 10
The willingness to download and purchase excluding Beautiful–Ugly and Good–Bad (stepwise).

Step #	Beta Download ($R^2 = 0.378$)	p	Step #	Beta Purchase ($R^2 = 0.304$)	p
1 Expensive–Cheap	–0.179**	0.000	1 Expensive–Cheap	–0.153**	0.000
2 Professional–Unprofessional	–0.210**	0.000	2 Professional–Unprofessional	–0.197**	0.000
3 Relaxed–Stiff	–0.080**	0.000	3 Unique–Ordinary	–0.137**	0.000
4 Unique–Ordinary	–0.125**	0.000	4 Relaxed–Stiff	–0.095**	0.000
5 Calm–Exciting	0.119**	0.000	5 Realistic–Unrealistic	–0.099**	0.000
6 Realistic–Unrealistic	–0.100**	0.000	6 Strong–Weak	–0.104**	0.000
7 Quiet–Loud	–0.117**	0.000	7 Hard–Soft	0.076**	0.001
8 Passive–Active	0.080**	.001	8 Slow–Fast	0.086**	0.000
9 Strong–Weak	–0.106**	0.000	9 Quiet–Loud	–0.116**	0.000
10 Hard–Soft	0.059**	.007	10 Calm–Exciting	0.082**	0.001
11 Complex–Simple	0.054**	.003	11 Concrete–Abstract	0.042*	0.038
12 Happy–Sad	–0.055*	.011			
13 Concrete–Abstract	0.043*	.027			
14 Old–Young	0.042*	.030			

* $p < 0.05$.
** $p < 0.01$.

unique, exciting, young, masculine, relaxed, strong, soft and three-dimensional. Without the adjective pairs Beautiful–Ugly and Good–Bad, the following adjectives also revealed to be positive predictors of the overall evaluation (Table 9): *happy, expensive, quiet, realistic, colorful, and simple*.

The following ends of the semantic differentials positively predicted the willingness to click app icons (Table 7): *good, beautiful, unique, active, colorless, exciting, quiet, masculine, two-dimensional, realistic and soft*. Without the adjective pairs Beautiful–Ugly and Good–Bad, the following adjectives also revealed to be positive predictors of the willingness to click (Table 9): *expensive, professional, happy, strong, relaxed, and simple*.

The following ends of the semantic differentials positively predicted the willingness to download the imagined app that the icon belongs to (Table 8): *good, beautiful, unique, exciting, professional, quiet, active, realistic, abstract, soft and masculine*. Without the adjective pairs Beautiful–Ugly and Good–Bad, the following adjectives also revealed to be positive predictors of the willingness to download (Table 10): *expensive, relaxed, strong, simple, happy, and young*.

The following ends of the semantic differentials positively predicted the willingness to purchase the imagined app that the icon belongs to (Table 8): *good, beautiful, unique, sad, fast, realistic, soft, professional and colorless*. Without the adjective pairs Beautiful–Ugly and Good–Bad, the following adjectives also revealed to be positive predictors of the willingness to purchase (Table 10): *expensive, relaxed, strong, quiet, exciting, and abstract*.

Lastly, we ran the MLRA models with the variable on how many mobile games participants download per week as a control variable to investigate systematic effect on rating. Section 3.1 (Table 1) stated that the majority of participants (75.4%) did not download any game apps on a weekly basis. Including this variable in the analyses did not alter the ratings in a significant manner. The number of game apps downloaded had a statistically significant effect ($\beta = -0.034, p < 0.01$) in the overall evaluation of the icon, but as the effect is quite small, it can be considered irrelevant in these results.

5. Discussion

Using semantic differential scale (22 adjective pairs), this study investigated the relationship between consumer perceptions of app icons and icon successfulness, measured by 1) overall evaluation of the icon, 2) willingness to click the icon, 3) willingness to download the imagined app and, 4) willingness to purchase the app. The study design was a vignette study, in which participants ($n = 569$) were assigned to evaluate 4 randomized icons from a total of pre-selected 68 game app icons across 4 categories (concrete, abstract, character and text). This

resulted in a total of 2276 individual icon evaluations. The goal was to discover aesthetic qualities that are likely to predict consumer behavior related to clicking on app icons as well as downloading and purchasing apps on app stores.

A clear pattern was displayed by the ratings of the MLRA including all 22 adjective pairs (Table 4) in that the likelihood to icon successfulness can be predicted by the following set of adjectives: *beautiful* (vs. *ugly*), *good* (vs. *bad*), and *unique* (vs. *ordinary*). Icons that were associated with these adjectives projected a positive overall evaluation and willingness to click the app icon as well as download and purchase the imagined app. The polar opposite of these adjectives on the semantic scale has an equally negative effect on the aspects of icon successfulness.

The adjectives “beautiful” and “good” were statistically significant in all cases, which was to be expected. As the adjectives are of general nature, they may reflect more of an overall estimate of aesthetic quality of an icon which poses a potential limitation that should be considered in future studies. Adjective pairs related to aesthetic qualities, such as Colorful–Colorless, Realistic–Unrealistic and Two-dimensional–Three-dimensional, are perhaps more specific in nature and thus express more variation in the ratings seen on Table 4. Nevertheless, this insight is valuable as it contrasts prior results (Shaikh, 2009) on onscreen typeface design and usage that was implemented as a basis for our experiment. Shaikh’s (2009) results indicate that not all typefaces for online content should convey beauty, particularly if it is not consistent with the meaning and context of the text. The findings in this study indicate that in app icon successfulness, beauty is an important factor in all cases regardless of the context.

Prior research emphasizes the importance of icon uniqueness related to task performance and user preference (Arend et al., 1987; Creusen and Schoormans, 2005; Creusen et al., 2010; Dewar, 1999; Goonetilleke et al., 2001; Huang et al., 2002). The results in this study support this notion as the adjective “unique” is statistically significant in each of the four variables in the MLRA including all 22 adjective pairs (Table 4) along with the post-hoc MLRA that were performed without the adjective pairs Beautiful–Ugly and Good–Bad (Table 6). Icon memorability is suggested as one of the key design elements for app icons by the developer guides of leading app stores,^{6,7} which is likely due to the large mass of app icon material available for consumers on app stores. Uniqueness helps app icons stand out in a display

⁶ Apple Developers, “App icon,” <https://developer.apple.com/ios/human-interface-guidelines/icons-and-images/app-icon/> (Accessed January 30, 2018).

⁷ Android Developers, “Iconography,” <http://www.androiddocs.com/design/style/iconography.html> (Accessed January 30, 2018).

of many other offerings. Hence, on the basis of the ratings in our analyses (Tables 4 and 6), we suggest that app icons need to be distinguishable to successfully attract consumers. Evidently, icon uniqueness is a combination of multiple features. However notably, a comparison between the four icon categories (abstract, concrete, character and text) in this study indicate that abstract icons were perceived as unique more often than icons from the other categories. Thus, abstract elements may enhance perceived icon uniqueness.

The post-hoc MLRA that were performed without the adjective pairs Beautiful–Ugly and Good–Bad (Table 6) exposed other significant effects in addition to icon uniqueness, that may perhaps explicate icon successfulness on a more detailed level. The results in Table 6 indicate that the likelihood to a higher overall evaluation as well as clicking, downloading and purchasing can be predicted by the following adjectives: *professional* (vs. *unprofessional*), *expensive* (vs. *cheap*), *soft* (vs. *hard*), *strong* (vs. *weak*), *relaxed* (vs. *stiff*), *realistic* (vs. *unrealistic*), *exciting* (vs. *calm*) and *quiet* (vs. *loud*).

Product presentation has been shown to affect consumer perceptions, meaning that if the presentation conveys high quality, consumers also perceive the product to be of high quality, and vice versa in relationship to low quality presentation (Silayoi and Speece, 2004). This way, the representation can be favorable or unfavorable to the content. The evidence in this study suggests that this theory may apply to app icons, as both the adjectives “professional” and “expensive” convey high quality. From the pool of the 68 game app icons used in this study, these adjectives are associated with such aesthetic app icon qualities as e.g. rounded corners, use of color gradient, shading and highlighting.

Prior research (Burgers et al., 2016) has established a connection with positive consumer attitudes and the use of visual metaphors in app icons. In consumer research, the use of metaphors has been shown to enhance appreciation of a product, because it is much like solving a puzzle, which rewards the consumer and thus inspires positive evaluations (Phillips and McQuarrie, 2009). For example, the product attribute of softness can be metaphorically represented by feathers or kittens, whereas strength can be represented by an image of a lion (Fenko et al., 2018). The statistical significance of the adjectives “soft” and “strong” highlights these prior observations. In this study, “soft” which is the opposite of “hard”, is associated with such aesthetic app icon qualities as e.g. desaturated colors and objects depicted in the icon that are perceived as delicate, e.g. animal fur. On the other hand, “strong” which is the opposite of “weak”, is associated with bold colors and hard-rendered surfaces, as well as objects depicted in the icon that are perceived powerful, e.g. a flying saucer or a hammer. It is believed that positive emotion between consumer and product brings extra value and increases the possibility of purchase (Cho and Lee, 2005). Furthermore, positive impressions have been considered as an important part of consumer perception (Yun et al., 2003). The statistical significance of the adjectives “relaxed” and “quiet” emphasize this observation as they are emotionally engaging qualities that can be perceived positive. Prior results have shown that vivid, highly saturated colors are perceived as exciting by consumers (Labrecque and Milne, 2011). The adjective “exciting”, which is the opposite of “calm” in this study, supports these findings as the icons that received high ratings for the adjective “exciting” express bold colors, similar to the icons perceived as “strong”. The icons perceived as “exciting” also highly correlate with the stimulus depicted in the icon, e.g. the face of an angry dragon or riding a motorcycle.

Icon concreteness is the extent to which it depicts real objects (Isherwood et al., 2007), whereas icon abstractness tends to have less obvious connections with real objects (McDougall et al., 1999). In this study, the positive ratings for the adjective “realistic” may be correlated to icon concreteness, which has been proven a significant predictor in icon effectiveness in prior studies (Hou and Ho, 2013; McDougall and Reppa, 2008; McDougall et al., 1999, 2000; Moyes and Jordan, 1993; Rogers and Osborne, 1987). However, the adjective “concrete” was found to be of no significant effect in the results (Tables 4 and 6), which

justifies further research on the relationship of concreteness and realism within the genre of app icons.

The main observation of the results is not only the similarities that increase our insight into this topic, but also the differences in conjunction with the recurrence of statistically significant variables that may explicate consumer perceptions of app icons on a more detailed level. In spite of the findings in the MLRA that were performed without the adjective pairs Beautiful–Ugly and Good–Bad (Table 6), it is important to note that both “beautiful” and “good” are significant in predicting consumer interaction with app icons.

The findings in this study exposed gaps in prior icon design theories, which they did not replicate to the following extent. From the perspective of previous literature on effective icon design, the statistical insignificance of the adjective pairs Concrete–Abstract and Complex–Simple was unexpected. It has been widely speculated that the Concrete–Abstract (e.g. 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) and Complex–Simple (e.g. Choi and Lee, 2012; Goonetilke et al., 2001; McDougall and Reppa, 2008, 2013; McDougall et al., 2016) relationship predicts icon successfulness. However, the results of our experiment contrast this statement as neither of the adjective pairs was statistically significant in the first MLRA (Table 4). In the second MLRA (Table 6), the adjective pair Complex–Simple was only marginally statistically significant in two of the variables (i.e. overall evaluation and downloading). This calls for more research on app icons, as the Concrete–Abstract and Complex–Simple theories were initially established within other icon genres and have not yet been explored further in relationship to app icons.

























5.1. Practical implications

Before setting practical implications, it is essential to understand the scope of the empirical study and the scope of possible guidelines that can be drawn. The broad objective of the study was to increase the understanding of how people's aesthetic perceptions of graphical user interface elements affect people's willingness to interact with those elements. In this study, we selected game app icons as the operationalization and/or case of graphical user interface element, and self-reported overall evaluation as well as willingness to click (the icon), download and purchase the app related to the icon as operationalization of GUI element successfulness. It should be noted that this study did not measure or investigate the relationship between graphical features of the icons and aesthetic perception. The study investigated relationships between the aesthetic perception and willingness to interact. Therefore, the study is unable to directly or reliably inform how graphical user interface elements should be designed in terms of their features, rather it can inform what kinds of aesthetic perceptions graphical user interface elements (i.e. icons) should be brought to evoke. Hence, the recommendations related to graphical features herein are based on qualitative assessment of the mean scores of different adjectives in icon ratings. The results and guidelines are directly applicable to the context of mobile (game) app icons, and with some hesitation, all icons. The results could also be applied all the way up to discussing best practices related to any graphical user interface elements but naturally with increased caution as the external validity diminishes the more general the context in which the knowledge from the results is applied in. Naturally, more similar research is needed across categories of GUIs to enforce and enrich the normative knowledge surrounding the area. Practical implications directly following the empirical results of the study are listed in the following.

Design implication 1: First and foremost, the results unsurprisingly suggest that to increase consumer interaction in terms of app icon successfulness (i.e. overall evaluation, willingness to click an icon as well as download and purchase the imagined app behind the icon), the

Table 11

Top 6 icons with the highest score in overall evaluation, willingness to click the icon, as well as download and purchase the imagined app on a seven-point scale (1 = lowest and 7 = highest).

#	Overall evaluation		Willingness to click		Willingness to download		Willingness to purchase	
	Icon	Mean	Icon	Mean	Icon	Mean	Icon	Mean
1		4.77		4.22		4.00		3.68
2		4.52		4.21		3.95		3.63
3		4.51		4.19		3.85		3.58
4		4.50		4.09		3.81		3.50
5		4.31		4.05		3.77		3.48
6		4.24		3.98		3.77		3.40

app icon should be perceived as high quality which is indicated by the results where the following perceptions predicted app icon successfulness across the board (Tables 4 and 6): *beautiful, good, professional, and expensive*. All these adjectives can be associated with general high quality. If cursorily investigating the icons that score high on these perceptions, they seem to share some of the following features (Appendix A): transparent parts on the outer layers, color gradients, shading and highlighting as well as high graphical fidelity.

Design implication 2: Separately from perceptions related to high quality (implication 1), *uniqueness* was another strong predictor of icon successfulness (Tables 4 and 6). Therefore, a successful app icon should be unique and memorable to stand out from the app store masses where there is a lot of icon material present. If cursorily investigating the icons that score high or low on the *uniqueness–ordinariness* continuum, they seem to share some of the following features (Appendix A): 1) icons rated as *unique* more commonly featured asymmetric and abstract shapes, use of various ends of the color spectrum as well as elements rarely encountered in daily life (e.g. a voodoo doll); 2) icons rated as *ordinary* broadly portrayed concrete, static shapes as well as objects typical to daily life (e.g. a house or a book).

Design implication 3: Beyond all perceptions that predicted all other factors of icon successfulness, *sadness* ($\beta = 0.059, p < 0.05$) and *fastness* ($\beta = 0.066, p < 0.01$) weakly predicted willingness to purchase the imagined app behind the icon (Tables 4 and 6). If cursorily investigating the icons that score high on these factors, they seem to share some of the following features (Appendix A): 1) icons rated as *sad* were generally dominated by a desaturated or dark color palette (e.g. shades of grey or pure black), and they depicted elements that can be perceived as unpleasant; 2) icons rated as *fast* illustrated actions or objects that

are typically associated with being rapid (e.g. a motorcycle or an airplane). A related observation is that icons with high scores of perceptions for things that are generally regarded as positive do not necessarily lead to higher icon successfulness. The indication that sadness predicts the willingness to purchase an app behind the icon is one example of this. Moreover, high overall evaluation score does not automatically lead to a high score in the willingness to click the icon, nor in the willingness to download or purchase the imagined app. Thus, it can be argued that app icons should incorporate more than one of the design implications in order to enhance the likelihood to consumer interaction.

Purely as illustrative examples, Table 11 introduces icons with the highest scores in overall evaluation of the icon design, the willingness to click the icon, as well as the willingness to download and purchase the imagined app. However, we wish to note again that this study did not study the relationship between icon features and successfulness per se. Therefore, any relationship between icon feature and success should be regarded as background data augmenting the focus of the study that was on the relationship between perception and successfulness.

5.2. Limitations and future research

App icon design is a complex matter with room for further investigation. This study was one of the first attempts to understand consumer perceptions of app icon successfulness by utilizing game app icons as data collection material. Moreover, this study attempted to rule out non-significant adjectives to aid future research on this topic. Nevertheless, this research has several limitations.

As is natural to any study, some compromises have to be made with

regards to study design as it is impossible to include the entire relevant phenomenon in the scope of a single study. In this study, as having all possible icons or icon categories as stimulus material, we selected one larger domain of iconography. We decided to select game app icons as the stimulus material of the study as 1) mobile game app icons are internally a homogenous category of graphical elements in the sense that they all share the same size, same possible color space and thus should eliminate unforeseen variability, 2) they are familiar to people from before-hand and participants can more effortlessly imagine countering such icons in their normal life, 3) game icons exhibit perhaps more heterogeneity in possible styles compared with icons related to utilitarian software/apps, and therefore, game apps may afford greater external validity and/or generalizability across icons, 4) currently game apps represent a hugely timely phenomenon as game apps are the most popular app category globally by several statistics.^{8,9,10}

Icons of new game apps were chosen for this study to eliminate the chance of app and icon familiarity and thus, systematic bias. Hence, the set of icons in this study may not represent the icons of top grossing game apps which consumers more commonly face on app stores. Therefore, consumer perceptions may have been affected by the comparison of top grossing game apps and their icons. This should not be regarded solely a limitation, as the sample of icons in this study represents the majority of game icon material on app stores and may thus give a more realistic perspective on consumer perceptions. However, this might contribute to the fact that the icons used in this study received ratings that can be perceived negative, e.g. ugly and cheap. The sample in this study is a nonprobability convenience sample, therefore it is not necessarily representative of all app store users.

Participants were uninformed about the purpose of the apps behind the icons in the experiment, as this could affect the results. Knowledge of the app may pose a risk of bias in user perceptions, thus the choice was made to eliminate possible confounding variables influencing the main objective of the study. This was further controlled by selecting new game app icons for the experiment that were not widely known. Therefore, this study may be limited in terms of external validity in order to maximize internal validity.

The research question of this study was “How does aesthetic perception of an icon lead to icon successfulness?” (as measured via likelihood to click, download, purchase and rate it highly). Hence, also our measurement is focused on icon aesthetics. This study did not include other factors aside from aesthetic qualities that contribute to a consumer's willingness to interact with app icons. In this study, we measured participants' self-reported willingness to interact with the icons presented to them. Alternatively, (intended) behavior could have also been tested by having participants actively click or select icons, or in a field experiment in a real app store to track user behavior. Due to the limitations of the measurement related to the dependent variables used in this study, further investigations could pursue logging behavioral metrics to increase validity of the study. However, in many cases collecting both the perceptual data on UI aesthetics and user behavior simultaneously may prove difficult.

Future research could be expanded in several directions. For one, investigating the Concrete-Abstract and Simple-Complex relationship regarding app icons would be beneficial, because the findings in this study did not support prior literature to that extent. Game app icons were used in this study to maximize internal validity. This introduces a possibility for

⁸ Statista, “Most popular Apple App Store categories in September 2018, by share of available apps,” <https://www.statista.com/statistics/270291/popular-categories-in-the-app-store/> (Accessed February 25, 2019).

⁹ Newzoo, “The global games market reaches \$99.6 billion in 2016, mobile generating 37%,” <https://newzoo.com/insights/articles/global-games-market-reaches-99-6-billion-2016-mobile-generating-37/> (Accessed February 25, 2019).

¹⁰ Statista, “Mobile phone gaming penetration in the United States from 2011 to 2020,” <https://www.statista.com/statistics/234649/percentage-of-us-population-that-play-mobile-games/> (Accessed February 25, 2019).

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. Moreover, future research could map out how participants would describe the imagined app behind the icon to see how icon design affects perceptions on the purpose of the app. As this study employed a large-scale quantitative approach, it provides a broad overview of the relationship between consumer perceptions and app icons. To acquire further, in-depth understanding of the topic, a qualitative approach is recommended. The aim of this study was to explore aesthetic qualities that contribute to consumer perceptions of app icon successfulness, yet other possible factors aside from aesthetic qualities could also be studied in the future, e.g. the number and type of downloaded apps and their effect on perceptions of icon successfulness. Finally, differences in perceptions between cultures, genders and personalities would be an interesting approach for future research to aid designing icons that correspond to the needs of different users.

6. Conclusion

This study explored how consumer perception affects app icon successfulness from an aesthetic perspective. Aesthetic appeal is subjective, which is a probable cause for variations in the results. However, the findings show evidence of consensus that proves an empirical relationship on consumer perceptions and icon successfulness. As the data sample in this study is particularly extensive, the results may be regarded as a contribution to overall knowledge. Revealing this phenomenon may be a building block that eventually leads to further theoretical implications around this topic. Furthermore, the design guidelines in this study assist app designers, developers and marketers when creating a key branding asset for app stores. As such, this study has laid the groundwork for future research that aims to understand consumer perceptions of app icons in graphical user interfaces and online storefronts.

Acknowledgments

This work was supported by the Finnish Funding Agency for Technology and Innovation TEKES (40111/14, 40107/14 and 40009/16) and participating partners, as well as Satakunnan korkeakoulusäätiö (<http://www.korkeakoulusaaio.fi/etusivu>) and its collaborators.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ijhcs.2019.04.004.

References

- Arab, F., Malik, Y., Abdurazak, A., 2013. Evaluation of PhonAge: an adapted smartphone interface for elderly people. In: Proceedings of the 14th IFIP TC 13 International Conference. Springer, Cape Town, South Africa, pp. 547–554. https://doi.org/10.1007/978-3-642-40498-6_44.
- Arend, U., Muthig, K.P., Wandmacher, J., 1987. Evidence for global superiority in menu selection by icons. *Behav. Inf. Technol.* 6, 411–426. <https://doi.org/10.1080/01449298708901853>.
- Ares, G., Piqueras-Fiszman, B., Varela, P., Marco, R.M., López, A.M., Fiszman, S., 2011. Food labels: do consumers perceive what semiotics want to convey? *Food Qual. Pref.* 22, 689–698. <https://doi.org/10.1016/j.foodqual.2011.05.006>.
- Black, A., 2017. Icons as carriers of information. In: Black, A., Luna, P., Lund, O., Walker, S. (Eds.), *Information design: Research and Practice*. Routledge, London, pp. 253–269.
- Blankenberger, S., Hahn, K., 1991. Effects of icon design on human-computer interaction. *Int. J. Man-Mach. Stud.* 35, 363–377. [https://doi.org/10.1016/S0020-7373\(05\)80133-6](https://doi.org/10.1016/S0020-7373(05)80133-6).
- Burgers, C., Eden, A., de Jong, R., Buningh, S., 2016. Rousing reviews and instigative images: the impact of online reviews and visual design characteristics on app downloads. *Mob. Media Commun.* 4, 327–346. <https://doi.org/10.1080/15213269.2016.1182030>.
- Böhmer, M., Krüger, A., 2013. A study on icon arrangement by smartphone users. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, Paris, France, pp. 2137–2146. <https://doi.org/10.1145/2470654.2481294>.
- Chen, C.C., 2015. 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 529 Springer, Cham. https://doi.org/10.1007/978-3-319-21383-5_2.
- Cho, H.-S., Lee, J., 2005. Development of a macroscopic model on recent fashion trends on the basis of consumer emotion. *Int. J. Consum. Stud.* 29, 17–33. <https://doi.org/10.1111/j.1470-6431.2005.00370.x>.
- Choi, J.H., Lee, H.-J., 2012. Facets of simplicity for the smartphone interface: a structural model. *Int. J. Hum. Comput. Stud.* 70, 129–142. <https://doi.org/10.1016/j.ijhcs.2011.09.002>.
- Creusen, M.E.H., Schoormans, J.P.L., 2005. The different roles of product appearance in consumer choice. *J. Prod. Innov. Manage.* 22, 63–81. <https://doi.org/10.1111/j.0737-6782.2005.00103.x>.
- Creusen, M.E.H., Verzyer, R.W., Schoormans, J.P.L., 2010. Product value importance and consumer preference for visual complexity and symmetry. *Eur. J. Mark.* 44, 1437–1452. <https://doi.org/10.1108/03090561011062916>.
- Crossley, L., 2003. Building emotions in design. *Design J.* 6, 35–41. <https://doi.org/10.2752/146069203789355264>.
- Dewar, R., 1999. Design and evaluation of public information symbols. In: Zwaga, H.J.G., Boersma, T., Hoonhout, H.C.M. (Eds.), *Visual Information for Everyday Use*. Taylor & Francis, London, pp. 285–303.
- Fenko, A., Schifferstein, H.N.J., Hekkert, P., 2010. Shifts in sensory dominance between various stages of user-product interactions. *Ergonomics* 41, 34–40. <https://doi.org/10.1016/j.apergo.2009.03.007>.
- Fenko, A., de Vries, R., van Rompay, T., 2018. How strong is your coffee? The influence of visual metaphors and textual claims on consumers' flavor perception and product evaluation. *Front. Psychol.* 9, 53. <https://doi.org/10.3389/fpsyg.2018.00053>.
- Forsythe, A., Mulhern, G., Sawey, M., 2008. Confounds in pictorial sets: the role of complexity and familiarity in basic-level picture processing. *Behav. Res. Methods* 40, 116–129. <https://doi.org/10.3758/BRM.40.1.116>.
- García, M., Badre, A.N., Stasko, J.T., 1994. Development and validation of icons varying in their abstractness. *Interact. Comput.* 6, 191–211. [https://doi.org/10.1016/0953-5438\(94\)90024-8](https://doi.org/10.1016/0953-5438(94)90024-8).
- Gittins, D., 1986. Icon-based human-computer interaction. *Int. J. Man-Mach. Stud.* 24, 519–543. [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., 2001. Effects of training and representative characteristics in icon design. *Int. J. Hum. Comput. Stud.* 55, 741–760. <https://doi.org/10.1006/ijhc.2001.0501>.
- Hou, K.-C., Ho, C.-H., 2013. 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.
- Horton, W., 1994. *The Icon book: Visual Symbols For Computing Systems and Documentation*. John Wiley & Sons, New York.
- Horton, W., 1996. Designing icons and visual symbols. In: *Proceedings of the CHI 96 Conference on Human Factors in Computing Systems*. Vancouver, Canada. pp. 371–372. <https://doi.org/10.1145/257089.257378>.
- Huang, S.-M., Shieh, K.-K., Chi, C.-F., 2002. Factors affecting the design of computer icons. *Int. J. Ind. Ergon.* 29, 211–218. [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., 2007. Icon identification in context: the changing role of icon characteristics with user experience. *Hum. Fact.* 49, 465–476. <https://doi.org/10.1518/001872007x200102>.
- Labrecque, L., Milne, G., 2011. Exciting red and competent blue: the importance of color in marketing. *J. Acad. Mark. Sci.* 40, 711–727. <https://doi.org/10.1007/s11747-010-0245-y>.
- Lin, C.-H., Chen, M., 2018. The icon matters: how design instability affects download intention of mobile apps under prevention and promotion motivations. *Electron. Commer. Res.* <https://doi.org/10.1007/s10660-018-9297-8>.
- Lodding, K.N., 1983. Iconic interfacing. *IEEE Comput. Graph. Appl.* 3, 11–20. <https://doi.org/10.1109/MCG.1983.262982>.
- McDougall, S.J.P., Curry, M.B., de Bruijn, O., 1998. Understanding what makes icons effective: how subjective ratings can inform design. In: Hanson, M. (Ed.), *Contemporary Ergonomics*. Taylor & Francis, London, pp. 285–289.
- McDougall, S.J.P., Curry, M.B., de Bruijn, O., 1999. 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. <https://doi.org/10.3758/BF03200730>.
- McDougall, S.J.P., de Bruijn, O., Curry, M.B., 2000. Exploring the effects of icon characteristics on user performance: the role of icon concreteness, complexity, and distinctiveness. *J. Exp. Psychol. Appl.* 6, 291–306. <https://doi.org/10.1037/1076-898X.6.4.291>.
- McDougall, S.J.P., Reppa, I., 2008. 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. <https://doi.org/10.1177/154193120805201822>.
- McDougall, S.J.P., Reppa, I., 2013. 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. https://doi.org/10.1007/978-3-642-39232-0_62.
- McDougall, S.J.P., Reppa, I., Kulik, J., Taylor, A., 2016. What makes icons appealing? The role of processing fluency in predicting icon appeal in different task contexts. *Appl. Ergon.* 55, 156–172. <https://doi.org/10.1016/j.apergo.2016.02.006>.
- Montgomery, D.C., Peck, E.A., 1992. *Introduction to Linear Regression Analysis*, Second ed. John Wiley & Sons, New York.
- Moreira, Á.V.M., Filho, V.V., Ramalho, G.L., 2014. Understanding mobile game success: a study of features related to acquisition, retention and monetization. *J. Interact. Syst.* 5, 2–13.
- Moyes, J., Jordan, P.W., 1993. 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*. Cambridge University Society, Cambridge, pp. 49–59.
- Orth, U.R., Malkewitz, K., 2008. Holistic package design and consumer brand impressions. *J. Mark.* 72, 64–81. <https://doi.org/10.1509/jmk.72.3.64>.
- Orth, U.R., Malkewitz, K., 2009. Good from far but far from good: the effects of visual fluency on impressions of package design. *Adv. Consum. Res.* 36, 211–212.
- Overby, E., Sabyasachi, M., 2014. Physical and electronic wholesale markets: an empirical analysis of product sorting and market function. *J. Manage. Inf. Syst.* 31, 11–46. <https://doi.org/10.2753/MIS0742-1222310202>.
- Peirce, C.S., 1932. Volumes I and II: principles of philosophy and elements of logic. In: Hartshorne, C., Weiss, P. (Eds.), *Collected Papers of Charles Sanders Peirce*. Harvard University Press, London.
- Phillips, B.J., 1996. Defining trade characters and their role in American popular culture. *J. Pop. Cult.* 29, 143–158. <https://doi.org/10.1111/j.0022-3840.1996.1438797.x>.
- Phillips, B.J., McQuarrie, E.F., 2009. Impact of advertising metaphor on consumer belief: delineating the contribution of comparison versus deviation factors. *J. Advert.* 38, 49–62. <https://doi.org/10.2753/JOA0091-3367380104>.
- Reppa, I., McDougall, S.J.P., 2015. When the going gets tough the beautiful get going: aesthetic appeal facilitates task performance. *Psychol. Bull.* 22, 1243–1254. <https://doi.org/10.3758/s13423-014-0794-z>.
- Rogers, Y., Osborne, D.J., 1987. Pictorial communication of abstract verbs in relation to human-computer interaction. *Br. J. Psychol.* 78, 99–112. <https://doi.org/10.1111/j.2044-8295.1987.tb02229.x>.
- Rundh, B., 2009. Packaging design: creating competitive advantage with product packaging. *Br. Food J.* 111, 988–1002. <https://doi.org/10.1108/00070700910992880>.
- Salman, Y.B., Kim, Y., Cheng, H., 2010. 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.
- Schifferstein, H.N., Fenko, A., Desmet, P.M., Labbe, D., Martin, N., 2013. Influence of package design on the dynamics of multisensory and emotional food experience. *Food Qual. Prefer.* 27, 18–25. <https://doi.org/10.1016/j.foodqual.2012.06.003>.
- Schröder, S., Ziefle, M., 2008. Making a completely icon-based menu in mobile devices to become true: a user-centered design approach for its development. In: *Proceedings of the 10th international conference on Human computer interaction with mobile devices and services*. ACM, Amsterdam, the Netherlands, pp. 137–146.
- Shaikh, A.D., 2009. Know your typefaces! Semantic differential presentation of 40 onscreen typefaces. *Usab. N.* 11, 23–65.
- Silayoi, P., Speece, M., 2004. Packaging and purchase decisions: an exploratory study on the impact of involvement level and time pressure. *Br. Food J.* 106, 607–628. <https://doi.org/10.1108/00070700410553602>.
- Shu, W., Lin, C.-S., 2014. Icon design and game app adoption. In: *Proceedings of the 20th Americas Conference on Information Systems*. Georgia, USA.
- Underwood, R.L., Klein, N.M., Burke, R.R., 2001. Packaging communication: attentional effects of product imagery. *J. Prod. Brand Manage.* 10, 403–422. <https://doi.org/10.1108/10610420110410531>.
- van Rompay, T.J.L., Pruyn, A.T.H., Tieke, P., 2009. Symbolic meaning integration in design and its influence on product and brand evaluation. *Int. J. Des.* 3, 19–26.
- Wang, M., Li, X., 2017. Effects of the aesthetic design of icons on app downloads: evidence from an android market. *Electron. Commer. Res.* 17, 83–102. <https://doi.org/10.1007/s10660-016-9245-4>.
- Wiedenbeck, S., 1999. 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. <https://doi.org/10.1080/014492999191929>.
- Woolridge, D., Schneider, M., 2011. Your iOS app is your most powerful marketing tool. In: Woolridge, D., Schneider, M. (Eds.), *The Business of iPhone and iPad App Development*. Apress, Berkeley, pp. 61–96. https://doi.org/10.1007/978-1-4302-3301-5_4.
- Yun, M.H., Han, S.H., Hong, S.W., Kim, J., 2003. Incorporating user satisfaction into the look-and-feel of mobile phone design. *Ergonomics* 46, 1423–1440. <https://doi.org/10.1080/00140130310001610919>.

PUBLICATION IV

**Demographic factors have little effect on aesthetic perceptions of icons:
A study of mobile game icons**

Henrietta Jylhä & Juho Hamari (2021)

Internet Research (ahead-of-print)
DOI: 10.1108/INTR-07-2020-0368

Publication reprinted with the permission of the copyright holders.

Demographic factors have little effect on aesthetic perceptions of icons: a study of mobile game icons

Demographic effects in aesthetic GUI perceptions

Henrietta Jylhä and Juho Hamari

Gamification Group, Tampere University, Tampere, Finland

Abstract

Purpose – Customization by segmenting within human–computer interaction is an emerging phenomenon. Appealing graphical elements that cater to user needs are considered progressively important, as the way a graphic is visually represented can greatly contribute to the interaction. However, aesthetic perceptions are subjective and may differ by target group. Understanding variations in user perceptions may aid in design processes; therefore, we set out to investigate the effects of demographic differences relating to perceptions of graphical user interface (GUI) element (i.e. game app icon) aesthetics.

Design/methodology/approach – The authors employed a vignette experiment with random participant ($n = 513$) assignment to evaluate 4 icons from a total of 68 pre-selected mobile game icons using semantic differential scales. This resulted in a total of 2052 individual icon evaluations. Regression analyses were performed with the effects of age, gender and time using graphical user interfaces (i.e. app stores) and the interactions of these variables relating to perceptions of GUI element aesthetics.

Findings – The results indicate that, overall, demographic factors have relatively little effect on how icons are perceived. Significant relations suggest that experienced users, younger audiences and women are more critical in their perception of aesthetic excellence, and that perceptions change for younger women. The implications of the findings are discussed via adaptive decision-making theory.

Originality/value – In the context of graphical user interface element aesthetics, demographic differences have received minimal attention as moderating variables regardless of their relevance in design and development. Hence, it merits further research.

Keywords Iconography, Aesthetics, Demographics, User perception, Graphical user interface, Human–computer interaction

Paper type Research paper

1. Introduction

Demographic differences in designing aesthetically pleasing graphical user interface (GUI) elements have become prevalent due to increasing demands for customization within human–computer interaction (Norman, 2004; Tractinsky *et al.*, 2000). As a wide variety of daily communication is realized via user interfaces of different devices, designers are presented with new opportunities and challenges to create visually effective GUI elements for their targeted consumer group. Moreover, perceptions of successful (i.e. appealing) visual aesthetics are subjective (Zen and Vanderdonck, 2016), which complicates creating balanced user experiences for critical masses. Especially in mobile environments, the adoption of mobile game applications is a complex entity of varying perceptions, such as gender, content price and quality and time spent playing mobile games (Pappas *et al.*, 2019). Therefore, insight into what aspects of GUI element aesthetics are preferred by segmentation is needed.

User interfaces that adapt to individual preferences have been shown to lead to higher ratings in look and feel as well as long-term usage of platforms (Debevc *et al.*, 1996; Hartmann

Received 7 July 2020
Revised 7 October 2020
11 January 2021
22 June 2021
Accepted 22 June 2021

© Henrietta Jylhä and Juho Hamari. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at <http://creativecommons.org/licences/by/4.0/legalcode>

Academy of Finland Flagship Programme [337653 - Forest-Human-Machine Interplay (UNITE)].



Internet Research
Emerald Publishing Limited
1066-2243
DOI 10.1108/INTR-07-2020-0368

et al., 2007a; Sarsam and Al-Samarraie, 2018). Considering that ineffective interface usability tends to affect older age groups (Johnson and Finn, 2017) due to visual acuity changes (Huang, 2013), and that age is likely to contribute to users' skill level and experience with technology (KnowItAll Ninja, 2016), it can be considered a meaningful factor in GUI aesthetics and design which merits for further research on the topic.

Regarding gender differences in the field of human-computer interaction and visual aesthetics, the norm has been that preferences of male and female users differ to a significant degree (Genuine, 2013); however, new trends of more unisex patterns have been discovered (Morris *et al.*, 2005). In the future, offering gender-neutral options for user interfaces could be one solution to the possible minimizing of gender differences (Boiano *et al.*, 2006). Due to the change of the cultural atmosphere, there is a need to examine the effects of gender in this context.

Time interacting with interfaces contributes to impressions on aesthetics, nevertheless, this topic has received relatively little attention especially considering mobile interfaces (Miniukovich and De Angeli, 2014). Time affects several user attributes, preferences and expectations (Hartmann *et al.*, 2008; Thüring and Mahlke, 2007) that can lead to various outcomes concerning interface design. The norms of device interaction suggest that, grave alterations to GUI designs may hinder user adjustment and lead to frustration, and thus gradual changes are advised (KnowItAll Ninja, 2016). As the frequency of use is related to aesthetic perceptions on a general level, it is an important variable in determining the subjective experience.

Prior research has indicated that not only the main effects of age, gender and time are to be investigated, but also the interactions of these demographics should be taken into account, as significant relationships have been found between, e.g. age and gender on technology adoption (Morris *et al.*, 2005) as well as gender and time on mobile entertainment (Hsiao and Chen, 2016; Pappas *et al.*, 2019). Research regarding demographic differences in relation to aesthetic perceptions of GUI elements is scarce at present. The rapid progress of GUI design further justifies the current undertaking.

As described, different results exist on user interface aesthetics and the trends regarding age, gender and time spent interacting with devices, thus more work is needed to understand how the interplays of these particular demographics may offer a deeper understanding on perceptions of GUI aesthetics, and how they may affect further design and research processes. To address this gap, we observe user perceptions on GUI aesthetics based on adaptive decision-making theory (Payne *et al.*, 1993). Used in previous evaluations of interface quality (e.g. Hartmann *et al.*, 2007a, b, 2008), this approach allows interpreting the results with a conception that user judgment is adaptive and based on the task, context and background-experience. This theory is valid particularly in choice situations where no single alternative is best on all attributes (Beresford and Sloper, 2008).

The large-scale quantitative demographic data in this study was collected via a vignette experiment with random participant ($n = 513$) assignment, where the task was to evaluate 4 icons from a total of 68 pre-selected game app icons across 4 categories (concrete, abstract, character and text) using semantic scales. This resulted in a total of 2052 individual icon evaluations. Based on the results, our study presents insight into the effects of age, gender and time using graphical user interfaces (i.e. *app stores*) and the interactions of these variables relating to perceptions of GUI element aesthetics. Knowledge of these relations allows for theoretical and practical guidelines in the design process of personalized graphical user interface elements.

2. Background

2.1 Aesthetics perceptions in graphical user interfaces

Visual aesthetics in graphical user interface design can be defined as aesthetically pleasing or attractive computer-based environments, reflecting the format in which the content and services are presented as well as the design look and feel and overall experience with a system (Ahmed *et al.*, 2009; Hartmann *et al.*, 2007b; Jennings, 2000). As a research field, it focuses on the

user's subjective judgment on how aesthetic a system or a product is (Lee and Koubek, 2011), an increasingly important area in human–computer interaction due to the wide adaptation of devices for everyday actions. Aesthetics within human–computer interaction can be divided into classical and expressive aesthetics (Ahmed *et al.*, 2009; Hartmann *et al.*, 2008; Lavie and Tractinsky, 2004). Classical aesthetics refers to clear designs, whereas expressive aesthetics refer to more creative designs. Especially concerning interface icons, aesthetic appeal has been described as mild aesthetic experiences that refer to the power to attract users (McDougall *et al.*, 2016). In system design, the structure of information has been linked with perceived aesthetics as well as usability (Ahmed *et al.*, 2009; Cyr, 2009). Interaction with user interfaces is realized via graphical elements providing intuitiveness and immediate visual feedback, such as windows, menus and icons (Linux Information Project, 2004). Aesthetics in graphical user interface design has been proven an integral part of a positive user experience as well as user engagement (Kurosu and Kashimura, 1995; Ngo *et al.*, 2000; Overby and Sabyasachi, 2014; Salimun *et al.*, 2010; Tractinsky *et al.*, 2000). Positive user experience is important for successful human–computer interaction, as the user may abandon an interface that is related with a negative experience. User experience is connected to visual aesthetics to an increasing extent (Debevc *et al.*, 1996; Hartmann *et al.*, 2007a; Sarsam and Al-Samarraie, 2018); hence, an attractive user interface is important when aiming for successful human–computer interaction as well as positive commercial performance (Gait, 1985; Lin and Yeh, 2010).

Perceptions of effective visual aesthetics have been attempted to assess via various theories and tools (e.g. Choi and Lee, 2012; Hassenzahl *et al.*, 2003; Maity *et al.*, 2015; Ngo *et al.*, 2000; Ngo, 2001; Ngo *et al.*, 2003; Salimun *et al.*, 2010; Zen and Vanderdonckt, 2016), yet robust guidelines for designing GUI elements are lacking due to the complexity of the topic. Prior research (Maity *et al.*, 2015; Ngo *et al.*, 2000) has found correlations between metric-based aesthetic value and the aesthetics ratings of design experts, artists and users. However, these results were only partly supported by a similar study (Zen and Vanderdonckt, 2016). Another study (Salimun *et al.*, 2010) contrasted prior literature (Ngo, 2001; Ngo *et al.*, 2003) in that some metrics, such as symmetry and cohesion, influence results more than others. In addition to metric-based instruments, aesthetic value of graphical user interfaces has been measured by survey-based methods (Choi and Lee, 2012; Hassenzahl *et al.*, 2003; Jylhä and Hamari, 2020) aligned with user perceptions. Prior contradictory results in evaluation theories and tools of aesthetics in interface research are perhaps due to analyzing user interfaces as entities (Zen and Vanderdonckt, 2016). As user interfaces essentially consist of several elements with different purposes, it motivates investigating GUI elements separately rather than as an entity. Therefore, in this study, we scaled the sample into single interface components, i.e., icons. While icons do not constitute a GUI solitarily, icon-based interfaces are highly common at present. This justifies using icons as study material for evaluating the effect of demographic differences within user perceptions of GUI element aesthetics.

2.2 Demographic differences in interaction design

Prior literature on the effects of demographic differences in human–computer interaction and aesthetics suggests an impact on user perceptions, motivations and design processes (Creusen, 2010; Johnson and Finn, 2017; Oyibo *et al.*, 2016, 2018). However, in the context of graphical user interface element aesthetics, demographic differences have received minimal attention as moderating variables (Oyibo *et al.*, 2018). Advancing knowledge in the topic is beneficial to scholars and practitioners alike as contributions of this study may be adapted in further examining and designing user interface systems within the context of human–computer interaction.

Regarding interface design and age, it has been indicated that younger people tend to focus more on hedonic pleasure, whereas older people prefer a more utilitarian approach (Hsieh *et al.*, 2004; Johnson and Finn, 2017; Wallendorf and Arnould, 1988). Research has

shown that younger people are more critical towards aesthetics than older people, who were found to be indifferent about color schemes, while younger people were found to prefer moderate-temperature (green and orange) to extreme temperature (blue and red) color schemes (Oyibo *et al.*, 2018). Thus, interface designers are prompted to put effort in aesthetics considerations in order to appeal to younger audiences. Prior studies regarding age and technology have indicated a digital divide between generations in which younger age groups are less affected by social influence due to early technology adoption (Morris and Venkatesh, 2000; Venkatesh *et al.*, 2000). Moreover, older generations tend to experience more anxiety relating to human-computer interaction than younger people, as the process of adapting to new devices may be more time-consuming with age due to the diminishing of cognitive abilities such as memory capacity, symbol and language comprehension (Chung *et al.*, 2010; Creusen, 2010; Johnson and Finn, 2017; Rousseau *et al.*, 1998). Therefore, prior literature (Johnson and Finn, 2017) has suggested a number of design guidelines (e.g. the use of large fonts, maintaining visual consistency) in order to accommodate the aging population. Differences in perceptions between young and old users and the digital divide between age groups in today's society motivates for further observation of age in this context.

Concerning the effects of gender, implications have been made in terms of decision-making and information processing in that male users concentrate more on pragmatic aspects of technology, while female users are driven by social motivators (Sun and Zhang, 2006; Venkatesh and Morris, 2000). This means that in general, men are more orientated towards completing tasks and achievements than women (Hoffman, 1972; Minton and Schneider, 1980). On the other hand, women are more concerned with influential motivators and have been considered to be less likely to enjoy the use of information technology (Creusen, 2010; Hoffman, 1972; Venkatesh and Morris, 2000). Relating to interface design, males tend to prefer functional aspects (i.e. usability and symmetry), while females prefer expressive aspects (i.e. beauty and emotional value) (Creusen, 2010; Henry, 2002; Oyibo and Vassileva, 2017; Tuch *et al.*, 2010; Wallendorf and Arnould, 1988). In this regard, men can be considered more instrumental, whereas women are more symbolic and concerned with appearance. Furthermore, females have been found to be more sensitive to color and visual complexity in the context of user interfaces than males (Creusen, 2010; Reinecke and Gajos, 2014; Smith, 1995). A study in the context of mobile service adoption found no gender differences (Leong *et al.*, 2013). However, this has been countered by discovering that male and female users of mobile systems have different motivations, for example, males favor status and value and females prefer social and utilitarian orientations (Liu and Guo, 2017). This raises the need to examine the effect of gender especially in mobile environments.

In addition to age and gender, the effects of time using graphical interfaces (i.e. *app stores*) is to be taken into account. Prior literature has indicated that time affects user attributes, such as knowledge and skill level, as well as perceptions of system features, such as design and functionality (Lee and Koubek, 2011; Thüning and Mahlke, 2007). Frequent use of devices has shown to affect user preferences and expectations of visual aesthetics (Lee and Koubek, 2011). Moreover, users have been found to be selective with aesthetics based on experience (Hartmann *et al.*, 2008). Prior research on mobile entertainment has identified that time spent interacting with mobile systems affect user intentions and motivations, such as mobile game preference (Hsiao and Chen, 2016) and the level of investment on downloading mobile games (Pappas *et al.*, 2019). Involvement with GUI elements may impact users in several ways in regards to skill level, user experience, decision-making processes and perceptions of aesthetics. However, this topic has received relatively little attention especially considering mobile interfaces (Miniukovich and De Angeli, 2014). On the basis of prior literature and due to the lack of recent research, time is considered a valid factor in this study.

Demographic effects in the context of technology have shown to form multiple configurations of causal conditions. Prior research on technology adoption (Morris *et al.*, 2005)

has found a trend for more unisex pattern among younger people, suggesting that both younger women and men have received greater exposure to technology compared with the older generation, thus minimizing gender differences in this area. Furthermore, prior literature (Pappas *et al.*, 2019) has found a link between gender, content price and quality, as well as time spent playing mobile games: females who spend a lot of time playing games are more willing to overspend if the content is of high quality. This justifies the interpretation of effects between variables with respect to each other. Therefore, in addition to the independent variables age, gender and time, we employ the interaction effects of these variables. Interactions assess the relationship between an independent variable and dependent variable, moderated by a third variable (Aiken and West, 1991). This indicates that a third variable might influence the relationship between an independent and dependent variable, allowing for the observation of a more complex model where not only the main effects are studied. This can greatly expand understanding the relationships among variables in the model (Sweet and Grace-Martin, 2011).

2.3 Aesthetics, demographics and adaptive decision-making theory

In this paper, we observe user perceptions of GUI aesthetics in a theoretical framework of adaptive decision-making. The adaptive decision-making theory posits that an individual's use of decision strategies is an adaptive response of a limited-capacity information processor to the demands of complex decision tasks (Payne *et al.*, 1993). A person's repertoire of decision-making strategies depends on many factors, such as cognitive development, experience, and more formal training and education (Hartmann *et al.*, 2007a). As prior knowledge determines which strategies are available to a decision-maker, our elaboration of this theory hypothesizes that age, gender and experience are amongst the variables that affect decision-making behavior, contributing to individual differences.

The handful of studies that have investigated the relationship between interface aesthetics and demographics from the perspective of adaptive decision-making theory have consensus that the user's background plays an important role in the judgment of aesthetic appeal (Hartmann *et al.*, 2007a, b, 2008). These studies on aesthetics, usability and content relating to user interfaces theorize that preferences for user interface designs when the scenario of use is critical will be based on in-depth consideration, whereas for less serious scenarios, preferences will be based on selecting designs by general aesthetic impressions. The studies conclude that design priorities for aesthetics should be matched to user profile. Designers should not only know their audience, but also the audience's decision-making habits (i.e. preferences and expectations) that depend on interactions between decision-making criteria (e.g. design qualities such as content, aesthetics, functionality, usability).

As literature on this topic is limited, further investigation is justified. In the milieu of this theoretical framework and the study experiment, we expect users to evaluate the icon material by the strategy of desirability to the decision maker, a trait that is likely to be affected by demographic factors and background-experience. In particular, we hypothesize that the pattern of women being more drawn to the expressive aspects than men will continue, and that the time interacting with devices will have an effect on perceptions of visual aesthetics, i.e., the more time is spent, the more critical the users are towards the design aspects.

3. Methods and data

3.1 Participants

A nonprobability convenience sample was composed initially of 569 respondents who each assessed 4 game app icons through a survey-based within-subjects vignette experiment.

INTR

A within-subjects approach was chosen as opposed to between-subjects approach in order to expose each participant to all conditions (i.e. 4 icon evaluations by category) of the experiment. Due to insufficient representation, 15 responses without identifiable gender were removed. Additionally, 41 responses from older age groups were identified as outliers and removed, resulting in a total of 513 respondents with 2052 icon evaluations. Please refer to Table 1 for demographic details of participants.

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. The majority of participants resided in Finland (93.0%). The gender split across participants was rather equal, as only slightly more than half were male (52.4%). The mean age was 25.49 years (SD = 4.67 years; 16–39 years). As the majority of the respondents were from the same age group, the results of this study can be considered more representative of younger age groups. Most participants were university students (65.7%) and had a university-level education (41.1%). The majority of participants (40.2%) browsed app stores once per week. Most participants (75.6%) did not download any game apps on a weekly basis. Missing data (1.8%) was encountered for these two aforementioned items, as the frequency of app store usage and mobile game downloads were only asked from those who use a smartphone. To counter possible bias in the experiment, participants who did not download game apps frequently were instructed to answer based on their expectations of game app icons they

		<i>n</i>	%	
Gender	Male	269	52.4	
	Female	244	47.6	
Age by gender (SD = 4.67) (Mean = 25.49) (Median = 25.00)	–19	Male	14	2.7
		Female	8	1.6
	20–24	Total	22	4.3
		Male	105	20.5
		Female	120	23.4
	25–29	Total	225	43.9
		Male	103	20.1
		Female	75	14.6
	30–34	Total	178	34.7
		Male	27	5.3
Female		26	5.0	
35–39	Total	53	10.3	
	Male	20	3.9	
	Female	15	2.9	
Times browsing app stores	Total	35	6.8	
	0	147	28.7	
	1	206	40.2	
	2	74	14.4	
	3	34	6.6	
	4	10	1.9	
	5	12	2.3	
	More than 5	21	4.1	
	Missing	9	1.8	
	Game apps downloaded (per week)	0	388	75.6
1		94	18.3	
2		13	2.5	
3		7	1.4	
4		2	0.4	
Missing		9	1.8	

Table 1.
Demographic
information

might interact with. Two participants were randomly chosen and awarded a prize (Polar Loop 2 Activity Tracker). No other participation fees were paid. Participants were informed of the purpose of the study and assured anonymity.

3.2 Materials

Sixty-eight game app icons from Google Play Store were selected for the study. The decision to narrow down the sample to game app icons was made to eliminate further variability that might stem from the nature of the app and thus increase internal validity of the experiment, but also external validity in terms of results applied to the game icons. In order to avoid any systematic bias, 4 icons corresponding to dominant icon styles (concrete, abstract, character and text) were selected from each of 17 categories for game apps (action, adventure, arcade, board, card, casino, casual, educational, music, puzzle, racing, role playing, simulation, sports, strategy, trivia and word). Because icon design for app stores is category-dependent (Shu and Lin, 2014), we considered it justified to include icons from all categories. Prior literature highlights the relevance of concreteness and abstractness in icon design (e.g. 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; McDougall *et al.*, 2000; Moyes and Jordan, 1993; Rogers and Osborne, 1987); hence, they were included in this experiment. Looking at the icons on app stores, characters and typography are prevalent elements usually seen on app icons. It has been argued that faces on app icons are widely used because of the immediate impact and memorability they have due to neural processing of facial expressions (Chartboost, 2015). Furthermore, as the study design is based on prior research (Shaikh, 2009) on onscreen typeface and usage, text elements were included. During the selection phase we ensured that one icon from each category was dominantly characteristic of one of these 4 attributes.

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 as visually rich a sample of icons as possible, meaning that several different computer graphic techniques were included, such as 2D and 3D rendered images. The icons are presented in Table 2.

3.3 Measurements

Semantic differential scale was used to measure respondent evaluations of aesthetic aspects of the icons. A total of 22 adjective pairs was formulated and assigned to each icon. The polarity of the adjective pairs was reversed so that perceivably positive and negative adjectives did not align on the same side of the scale. All of the adjective pairs were chosen according to prior research (Shaikh, 2009) on onscreen typeface design and usage. Additionally, adjectives related to icons were added as suggested per previous literature on effective icon design. 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 were added to specifically measure the aesthetics of the icons include professional and unprofessional, colorful and colorless, realistic and unrealistic as well as two-dimensional and three-dimensional (Jylhä and Hamari, 2019).

INTR





































































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

Table 2.
Icons in the study

Developed further into a five-factor model entitled VISQUAL (Jylhä and Hamari, 2020), an instrument for measuring visual qualities of graphical user interface elements, the scale was used to observe underlying latent constructs in this study. VISQUAL consists of the aforementioned adjective pairs that were further divided into the following dimensions: *Excellence/Inferiority*, *Graciousness/Harshness*, *Idleness/Liveliness*, *Normalness/Bizarreness* and *Complexity/Simplicity*.

Table 3 lists the VISQUAL constructs and adjective pairs. Two versions of the model exist, the initial model with 22 adjective pairs and an adjusted model of 15 adjective pairs. In Table 3, the bolded adjective pairs represent those included in the adjusted model of 15 adjective pairs. Table 3 also presents an overview of the means and standard deviations. There were no outlier values and the range between the lowest and highest scores clustered closely to the average even though the 68 icons were quite different from each other. All the mean scores were between 3.5 and 4.5 for each evaluation. This indicates little skewness in the data.

Additional to the semantic scales, a seven-point Likert scale was utilized to measure the degree of disagree-agreement of the respondents with respect to the likelihood of them clicking, downloading, and purchasing the imagined app behind the icon with an instruction title: “Overall evaluation (judging by the icon alone)” followed by questions: “Compared to the mobile game icons I usually click, I would click this icon,” “Compared to the icons of mobile games I usually download, I would click this icon” and “Compared to the icons of mobile games I usually purchase, I would click this icon.” Respondents were provided the following options on the seven-point scale: “Strongly disagree,” “Disagree,” “Somewhat disagree,” “Neither agree nor disagree,” “Somewhat agree,” “Agree” and “Strongly agree.” Moreover, respondents were asked to give an overall evaluation score for the design of each icon by grading them on a seven-point scale to further assess consumer perceptions of icon successfulness.

Factor	Adjective pair	Mean	SD
Excellence/ Inferiority	<i>Good–Bad</i>	4.34	1.641
	<i>Professional–Unprofessional</i>	4.22	1.736
	<i>Beautiful–Ugly</i>	4.57	1.618
	<i>Expensive–Cheap</i>	4.83	1.563
	<i>Strong–Weak</i>	3.93	1.464
Graciousness/ Harshness	<i>Soft–Hard</i>	3.81	1.545
	<i>Relaxed–Stiff</i>	4.47	1.560
	<i>Masculine–Feminine</i>	4.34	1.388
	<i>Delicate–Rugged</i>	4.42	1.368
	<i>Happy–Sad</i>	3.80	1.507
	<i>Colorful–Colorless</i>	3.77	1.810
	<i>Warm–Cool</i>	3.97	1.436
Idleness/ Liveliness	<i>Fast–Slow</i>	3.87	1.576
	<i>Quiet–Loud</i>	4.12	1.601
	<i>Exciting–Calm</i>	3.96	1.452
	<i>Active–Passive</i>	3.97	1.708
Normalness Bizarreness	<i>Young–Old</i>	3.98	1.611
	<i>Concrete–Abstract</i>	4.03	1.998
	<i>Realistic–Unrealistic</i>	4.22	1.592
Complexity/ Simplicity	<i>Ordinary–Unique</i>	4.60	1.651
	<i>Simple–Complex</i>	4.69	1.669
	<i>Three-dimensional–Two-dimensional</i>	3.33	1.863

Table 3. Constructs in VISQUAL, means and standard deviations (adjusted 15 model items italics)

3.4 Procedure

The data was 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 was asked to evaluate app icons using seven-point 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 an initial “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. Last, demographic information (age, gender, etc.) was asked. The survey took about 10 minutes to complete. The survey was implemented via Surveygizmo, an online survey tool. All content was in English. The data was analyzed with IBM SPSS Statistics version 24 and Microsoft Office Excel 2016. The following section describes the results of the analysis.

4. Results

Regression analyses on the dependent variables (VISQUAL models and individual adjective pairs) were performed with age, gender and times browsing app stores (per week), as well as with the interaction terms of independent variables, namely age \times gender, age \times time, gender \times time and age \times gender \times time. In the analyses, the ratio-scale variable of age and time was used instead of the ordinal scales in Table 1. The independent variables age and time were centered prior to the analyses (Aiken and West, 1991), and the interaction terms were created from the prior centered variables. Prior to the analyses, multicollinearity test was performed on the independent variables as well as the interaction terms with variance inflation factors (VIF). No critical levels of multicollinearity were found between the variables. 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. First, regression analyses according to the VISQUAL model with 15 adjective pairs were performed. Please refer to Tables 4 and 5 for regression results.

When examining the results for statistically significant effects concerning age, gender and time, the results indicate that age ($\beta = -0.152, p = 0.033$) affects the *Excellence/Inferiority* dimension.

When observing these results, some statistically significant interactions between the independent variables were found. A two-way interaction between age and gender was found

for the dimension *Excellence/Inferiority* ($\beta = 0.170, p = 0.17$). Concerning age and time, a two-way interaction was found for the dimension *Normalness/Bizarreness* ($\beta = 0.167, p = 0.37$). No significant effect was found between gender and time or age, gender and time.

Second, regression analyses on the dependent variables according to the VISQUAL model with 22 adjective pairs were performed. Refer to Tables 6 and 7 for regression results.

Here, the results indicate that age ($\beta = -0.170, p = 0.17$) and time browsing app stores ($\beta = 0.173, p = 0.41$) affect the *Excellence/Inferiority* dimension.

Similar to the previous regression analyses (Table 5), a two-way interaction between age and gender was found for the dimension *Excellence/Inferiority* ($\beta = 0.193, p = 0.007$). Concerning age and time, a two-way interaction was found for the dimension *Normalness/Bizarreness*. ($\beta = 0.208, p = 0.009$). Additionally, a three-way interaction was found for the dimension *Normalness/Bizarreness* for age, gender and time ($\beta = -0.182, p = 0.025$). No significant effect was found between gender and time.

Lastly, regression analyses on the dependent variables as the individual 22 adjective pairs were performed. Refer to Tables 8 and 9 for regression results (organized per significant effects for further clarification).

When examining the results for statistically significant effects concerning age, gender and time, the adjective pairs *expensive-cheap* ($\beta = -0.163, p = 0.022$), *strong-weak* ($\beta = -0.172, p = 0.015$), *professional-unprofessional* ($\beta = -0.164, p = 0.022$) and *soft-hard* ($\beta = 0.157, p = 0.027$) were predicted by age. Regarding gender, *soft-hard* ($\beta = 0.058, p = 0.10$) and *warm-cool* ($\beta = 0.053, p = 0.019$) were predicted by gender. Finally, the adjective pair *ordinary-unique* ($\beta = -0.175, p = 0.039$) was affected by time spent browsing app stores on a weekly basis.

When examining these results, several statistically significant interactions between the independent variables were found. A two-way interaction between age and gender was found for the adjective pairs *expensive-cheap* ($\beta = 0.177, p = 0.013$), *strong-weak* ($\beta = 0.203, p = 0.004$), *professional-unprofessional* ($\beta = 0.181, p = 0.011$), *ordinary-unique* ($\beta = -0.156, p = 0.028$) and *good-bad* ($\beta = 0.147, p = 0.039$). Concerning age and time, a two-way interaction was found for the adjective pair *concrete-abstract* ($\beta = 0.170, p = 0.033$). A three-way interaction between age, gender and time was found for the adjective pairs *concrete-abstract* ($\beta = -0.166, p = 0.040$) and *exciting-calm* ($\beta = -0.162, p = 0.046$). No significant effect between gender and time was found.

5. Discussion

This study investigated the effects of age, gender and time using graphical user interfaces (i.e. *app stores*) relating to perceptions of GUI element (i.e. *game app icon*) aesthetics. The results

	Beta	<i>t</i>	<i>p</i>	Beta	<i>t</i>	<i>p</i>	Beta	<i>t</i>	<i>p</i>
	Age			Gender ^a			Time		
Excellence–Inferiority	<i>-0.152*</i>	<i>-2.134</i>	<i>0.033</i>	0.008	0.346	0.729	0.159	1.877	0.061
Graciousness–Harshness	0.045	0.635	0.525	-0.011	-0.496	0.620	0.044	0.513	0.608
Idleness–Liveliness	-0.064	-0.899	0.369	-0.018	-0.790	0.430	0.075	0.883	0.377
Normalness–Bizarreness	-0.014	-0.202	0.840	-0.030	-1.331	0.183	0.112	1.321	0.187
Complexity–Simplicity	-0.020	-0.933	0.351	0.024	1.125	0.261	0.032	1.435	0.151

Note(s): * = $p < 0.05$, ** = $p < 0.01$, statistically significant effects italicized. ^aFemales were coded with the higher variable value

Table 4. Regression analyses (15 items) adjusted model with age, gender and time spent browsing app stores (per week)

Table 5.
Regression analyses
(15 items) with Age \times
Gender, Age \times Time,
Gender \times Time, and
Age \times Gender \times Time

	Beta	<i>t</i>	<i>p</i>	Beta	<i>t</i>	<i>p</i>	Beta	<i>t</i>	<i>p</i>	Beta	<i>t</i>	<i>p</i>
		Age \times Gender			Age \times Time			Gender \times Time			Age \times Gender \times Time	
Excellence–Inferiority	<i>0.170*</i>	2.391	<i>0.017</i>	0.024	0.302	0.762	-0.129	-1.508	0.132	0.006	0.077	0.938
Graciousness–Harshness	-0.015	-0.209	0.835	-0.035	-0.438	0.661	0.052	0.639	0.856	0.052	0.639	0.523
Idleness–Livelihood	0.080	1.121	0.262	0.138	1.733	0.083	-0.126	-1.551	0.186	-0.126	-1.551	0.121
Normalness–Bizarreness	0.017	0.239	0.811	<i>0.167*</i>	<i>2.090</i>	<i>0.037</i>	-0.128	-1.488	0.137	-0.153	-1.890	0.059
Complexity–Simplicity	-0.033	-1.521	0.128	-0.023	-0.866	0.386	-0.034	-1.536	0.125	-0.018	-0.653	0.514

Note(s): * = $p \leq 0.05$, ** = $p < 0.01$, statistically significant effects italicized. ^aFemales were coded with the higher variable value

indicate that, overall, demographic factors have relatively little effect on how icons are perceived.

Observing the effects concerning age, gender and time with the VISQUAL models as dependent variables (Tables 4 and 6), statistically significant effects were found for age and time within the Excellence/Inferiority dimension. The negative correlation regarding age implies that the older the user, the more excellent (i.e. *good, professional, beautiful, expensive, and strong*) the icons were perceived, and the younger the user, the more inferior (i.e. *bad, unprofessional, ugly, cheap and weak*) the icons were perceived. This finding supports prior literature where younger audiences were found to be more critical towards GUI aesthetics than older audiences (Oyibo *et al.*, 2018). This might be explained by the notion that younger people tend to focus more on hedonic pleasure than older people (Hsieh *et al.*, 2004; Wallendorf and Arnould, 1988). The positive correlation concerning time suggests that the more time the user spends interacting with the interface, the less appealing the icons were rated. With the increase of time, users will naturally adapt to icon aesthetics that essentially repeat similar patterns, which may lead to developing a critical eye towards graphical elements. This way the users establish a taste for iconography over time, which might make users more selective.

Interaction effects were found between age and gender, age and time, as well as age, gender and time for the VISQUAL models (Tables 5 and 7) on the dimensions Excellence/Inferiority and Normalness/Bizarreness. On the basis of the positive correlation between age and gender, especially younger male users perceived the icons as excellent, and especially older female users perceived the icons as inferior. This finding is similar to prior literature in the way that women have been shown to appreciate aesthetics more than men, and might thus be more critical towards design aspects (Creusen, 2010; Oyibo and Vassileva, 2017). Likewise, early technology adoption within younger age groups and especially men (Morris and Venkatesh, 2000; Venkatesh *et al.*, 2000) might lead to better ratings, as they are perhaps generally more used to viewing game app icons. The positive interaction between age and time on the Normalness/Bizarreness dimension suggests that the perception of normalness (i.e. *concrete, realistic, ordinary*) tends to increase with time spent interacting with interfaces. The negative three-way interaction between age, gender and time on the same dimension further indicates that especially younger women evaluated icons as more normal when more time was spent using app stores. This suggests that icon aesthetics might be difficult to grasp in the beginning, which eventually changes as the user continues interacting with the interface. As noted previously, users tend to adapt to icon aesthetics thus losing some of their perceived uniqueness.

When examining the results concerning age, gender and time with individual adjective pairs as dependent variables (Table 8), age affected *expensive–cheap, strong–weak*,

	Beta	<i>t</i>	<i>p</i>	Beta	<i>t</i>	<i>p</i>	Beta	<i>t</i>	<i>p</i>
	Age			Gender ^a			Time		
Excellence–Inferiority	<i>-0.170*</i>	<i>-2.393</i>	<i>0.017</i>	0.002	0.069	0.945	<i>0.173*</i>	<i>2.046</i>	<i>0.041</i>
Graciousness–Harshness	0.071	1.001	0.317	0.023	1.004	0.316	0.035	0.407	0.684
Idleness–Liveliness	-0.044	-0.616	0.538	-0.030	-1.309	0.191	0.107	1.256	0.209
Normalness–Bizarreness	0.037	0.520	0.603	-0.020	-0.908	0.364	0.013	0.155	0.877
Complexity–Simplicity	0.067	0.946	0.344	0.029	1.297	0.195	0.105	1.241	0.215

Note(s): * = $p < 0.05$, ** = $p < 0.01$, statistically significant effects italicized. ^aFemales were coded with the higher variable value

Table 6. Regression analyses (22 items) with age, gender and time spent browsing app stores (per week)

Table 7.
Regression analyses
(22 items) with Age \times
Gender, Age \times Time,
Gender \times Time, and
Age \times Gender \times Time

	Beta	<i>t</i>	<i>p</i>	Beta	<i>t</i>	<i>p</i>	Beta	<i>t</i>	<i>p</i>	Beta	<i>t</i>	<i>p</i>
	Age \times Gender			Age \times Time			Gender \times Time			Age \times Gender \times Time		
Excellence–Inferiority	<i>0.193**</i>	2.713	<i>0.007</i>	0.026	0.329	0.742	-0.137	-1.602	0.109	0.006	0.073	0.942
Graciousness–Harshness	-0.053	-0.744	0.457	0.014	0.170	0.865	-0.038	-0.444	0.657	-0.002	-0.019	0.985
Idleness–Liveliness	0.068	0.949	0.343	0.102	1.276	0.202	-0.084	-1.041	0.150	-0.084	-1.041	0.298
Normalness–Bizarreness	-0.057	-0.798	0.425	<i>0.208**</i>	<i>2.604</i>	<i>0.009</i>	-0.048	-0.556	0.578	<i>-0.182*</i>	<i>-2.248</i>	<i>0.025</i>
Complexity–Simplicity	-0.065	-0.912	0.362	-0.074	-0.925	0.355	-0.060	-0.693	0.489	0.072	0.887	0.375

Note(s): * = $p \leq 0.05$, ** = $p < 0.01$, statistically significant effects italicized. ^aFemales were coded with the higher variable value

	Beta	<i>t</i>	<i>p</i>	Beta	<i>t</i>	<i>p</i>	Beta	<i>t</i>	<i>p</i>	Beta	<i>t</i>	<i>p</i>
		Age			Gender ^a			Time			Time	
Expensive-Cheap	<i>-0.163*</i>	-2.288	<i>0.022</i>	0.002	0.090	0.928	0.119	1.404	0.161			
Strong-Weak	<i>-0.172*</i>	-2.423	<i>0.015</i>	-0.024	-1.046	0.296	0.158	1.863	0.063			
Professional-Unprofessional	<i>-0.164*</i>	-2.300	<i>0.022</i>	0.018	0.817	0.414	0.149	1.754	0.080			
Soft-Hard	<i>0.157*</i>	2.209	<i>0.027</i>	<i>0.058*</i>	2.585	<i>0.010</i>	-0.084	-0.988	0.323			
Warm-Cool	0.053	0.739	0.460	<i>0.053*</i>	2.345	<i>0.019</i>	0.031	0.367	0.714			
Ordinary-Unique	0.108	1.518	0.129	0.009	0.420	0.675	<i>-0.175*</i>	-2.067	<i>0.039</i>			
Simple-Complex	0.116	1.633	0.103	0.026	1.150	0.250	0.037	0.432	0.666			
Young-Old	0.018	0.246	0.806	-0.041	-1.817	0.069	0.126	1.481	0.139			
Beautiful-Ugly	-0.083	-1.169	0.242	0.019	0.847	0.397	0.147	1.738	0.082			
Good-Bad	-0.120	-1.688	0.092	-0.013	-0.574	0.566	0.140	1.650	0.099			
Realistic-Unrealistic	-0.037	-0.521	0.603	-0.019	-0.852	0.395	0.119	1.405	0.160			
Concrete-Abstract	0.008	0.109	0.913	-0.030	-1.342	0.180	0.075	2.130	0.377			
Masculine-Feminine	-0.113	-1.591	0.112	-0.014	-0.602	0.547	0.064	0.757	0.449			
Delicate-Rugged	0.059	0.824	0.410	0.022	0.978	0.328	0.020	0.233	0.816			
Relaxed-Stiff	0.079	1.108	0.268	0.022	0.953	0.341	0.006	0.075	0.940			
Colorful-Colorless	-0.074	-1.034	0.301	-0.022	-0.987	0.324	0.059	0.700	0.484			
Three-Two-dimensional	-0.031	-0.436	0.663	0.008	0.376	0.707	0.081	0.959	0.338			
Active-Passive	-0.117	-1.647	0.100	-0.010	-0.439	0.661	0.062	0.734	0.463			
Fast-Slow	-0.043	-0.610	0.542	-0.026	-1.136	0.256	0.079	0.927	0.354			
Happy-Sad	0.127	1.781	0.075	-0.025	-1.107	0.268	0.036	0.430	0.667			
Exciting-Calm	-0.042	-0.594	0.553	-0.029	-1.295	0.196	0.048	0.569	0.570			
Quiet-Loud	0.072	1.016	0.310	0.025	1.114	0.266	-0.031	-0.367	0.713			

Note(s): * = $p < 0.05$, ** = $p < 0.01$, statistically significant effects italicized. ^aFemales were coded with the higher variable value

Demographic effects in aesthetic GUI perceptions

Table 8. Regression analyses with age, gender and time spent browsing app stores (per week)

Table 9.
Regression analyses with Age × Gender, Age × Time, Gender × Time, and Age × Gender × Time

	Beta	<i>t</i>	<i>p</i>	Beta	<i>t</i>	<i>p</i>	Beta	<i>t</i>	<i>p</i>	Beta	<i>t</i>	<i>p</i>
	Age × Gender			Age × Time			Gender × Time			Age × Gender × Time		
Expensive-Cheap	<i>0.177*</i>	2.488	<i>0.013</i>	-0.025	-0.316	0.752	-0.088	-1.020	0.308	0.053	0.655	0.512
Strong-Weak	<i>0.203**</i>	2.857	<i>0.004</i>	0.024	0.299	0.765	-0.122	-1.311	0.190	0.002	0.024	0.981
Professional-Unprofessional	<i>0.181*</i>	2.545	<i>0.011</i>	-0.027	-0.339	0.735	-0.100	-1.166	0.244	0.056	0.686	0.493
Ordinary-Unique	<i>-0.156*</i>	-2.197	<i>0.028</i>	0.155	1.941	0.052	0.127	1.485	0.138	-0.123	-1.520	0.129
Good-Bad	<i>0.147*</i>	2.062	<i>0.039</i>	0.036	0.453	0.651	-0.130	-1.511	0.131	-0.016	-0.198	0.843
Concrete-Abstract	0.019	0.264	0.792	<i>0.170*</i>	<i>2.130</i>	<i>0.033</i>	-0.108	-1.252	0.211	<i>-0.166*</i>	<i>-2.050</i>	<i>0.040</i>
Exciting-Calm	0.070	0.979	0.328	0.140	1.758	0.079	-0.080	-0.925	0.355	<i>-0.162*</i>	<i>-1.994</i>	<i>0.046</i>
Colorful-Colorless	0.104	1.464	0.143	-0.094	-1.175	0.240	-0.031	-0.356	0.722	0.102	1.252	0.211
Beautiful-Ugly	0.089	1.245	0.213	0.102	1.277	0.202	-0.135	-1.571	0.116	-0.072	-0.892	0.272
Simple-Complex	-0.086	-1.212	0.226	-0.066	-0.825	0.409	-0.018	-0.204	0.838	0.060	0.744	0.457
Realistic-Unrealistic	0.009	0.124	0.901	0.104	1.310	0.190	-0.108	-1.262	0.207	-0.083	-1.029	0.304
Young-Old	0.012	0.172	0.863	-0.019	-0.240	0.810	-0.091	-1.054	0.292	0.039	0.481	0.631
Active-Passive	0.111	1.562	0.118	0.088	1.101	0.271	-0.093	-1.078	0.281	-0.071	-0.869	0.385
Fast-Slow	0.048	0.679	0.497	0.066	0.826	0.409	-0.101	-1.175	0.240	-0.066	-0.808	0.419
Warm-Cool	-0.059	-0.830	0.407	0.043	0.535	0.593	-0.060	-0.700	0.484	-0.035	-0.430	0.667
Happy-Sad	-0.055	-0.771	0.441	-0.007	-0.086	0.931	-0.017	-0.201	0.841	0.025	0.309	0.758
Soft-Hard	-0.168	-2.360	0.323	0.064	0.807	0.420	0.024	0.283	0.777	-0.055	-0.674	0.500
Three-Two-dimensional	0.007	0.096	0.923	-0.021	-0.264	0.792	-0.049	-0.568	0.570	0.024	0.296	0.767
Masculine-Feminine	0.126	1.763	0.078	0.133	1.668	0.095	-0.073	-0.851	0.395	-0.121	-1.487	0.137
Delicate-Rugged	-0.061	-0.855	0.392	-0.094	-1.173	0.241	-0.009	-0.106	0.916	0.065	0.797	0.425
Relaxed-Stiff	-0.107	-1.496	0.135	0.027	0.339	0.735	0.013	0.146	0.884	-0.011	-0.130	0.897
Quiet-Loud	-0.063	-0.879	0.379	0.003	0.035	0.972	0.033	0.384	0.701	0.024	0.296	0.767

Note(s): * = $p \leq 0.05$, ** = $p < 0.01$, statistically significant effects italicized. ^aFemales were coded with the higher variable value

professional-unprofessional and *soft-hard*. The negative correlation for *expensive-cheap*, *strong-weak* and *professional-unprofessional*, as well as the positive correlation for *soft-hard*, indicates that the younger the user, the cheaper, weaker, unprofessional and harder the icons seemed. This strengthens the finding that young people are critical towards GUI aesthetics and perhaps more used to seeing app icons in general, leading to relatively poor evaluations. Gender differences were found for the adjective pairs *soft-hard* and *warm-cool*, indicating that female users perceived icons as harder and cooler compared to male users. These findings show that icon aesthetics are in certain ways perceived more harshly by women than men, perhaps relating to prior findings of women shown to be more sensitive to visual complexity (Creusen, 2010; Reinecke and Gajos, 2014; Smith, 1995). Time spent browsing app stores affected the adjective pair *ordinary-unique*. Hence, the longer the users spend time browsing app stores, the more ordinary the icons were perceived. This supports the previous indication of users developing a critical eye over time.

Several interaction effects with individual adjective pairs as dependent variables (Table 9) were found. Concerning age and gender, the adjective pairs *expensive-cheap*, *strong-weak*, *professional-unprofessional*, *ordinary-unique* and *good-bad* were statistically significant. These findings suggest that younger men perceived the icons as ordinary, while older women perceived the icons as cheap, weak, unprofessional and bad. This, again, may refer to the prior findings of women shown to appreciate aesthetics more than men as well as men being more comfortable with technology and perhaps generally more used to viewing game app icons. Furthermore, concerning age and time, it was found that for younger users, the more time they spend browsing app stores, the more concrete the icons seem. Finally, the negative three-way interaction between age, gender and time for the adjective pairs *concrete-abstract* and *exciting-calm* indicates that younger women evaluated icons as more concrete and exciting when more time was spent using app stores. These results strengthen prior findings of female users preferring expressive aspects (i.e. beauty and emotional value) (Creusen, 2010; Henry, 2002; Oyibo and Vassileva, 2017; Tuch *et al.*, 2010; Wallendorf and Arnould, 1988).

5.1 Theoretical contributions

The growing need for adaptive and appealing user interfaces requires more work in understanding how perceptions and demographic factors affect user interface design. This study adds to the topic of interaction research, where usability has dominated research partly at the expense of aesthetic considerations (Tractinsky *et al.*, 2000).

The results of this study contradict prior research on demographic factors and interface systems (Leong *et al.*, 2013) in that gender differences do exist, although they seem to be minimizing among the younger generation. The variety in perceptions between genders can partly be explained by strategies according to the adaptive decision-making theory, where users choose designs by filtering choices based on subjective impressions of aesthetics. As hypothesized previously, the decision-maker's strategies are dependent on the individual's history, which contributes to differences in perceptions.

The findings of this study are consistent with prior literature in that younger audiences are somewhat critical towards GUI aesthetics (Creusen, 2010; Hsieh *et al.*, 2004; Morris and Venkatesh, 2000; Oyibo and Vassileva, 2017; Oyibo *et al.*, 2018; Venkatesh *et al.*, 2000; Wallendorf and Arnould, 1988). However, interestingly, as the sample in this study focused on younger age groups and no significant effect was found between gender and time using interfaces, it seems that the unisex pattern identified by prior research (Morris *et al.*, 2005) is continued: gender differences are not as visible among younger users as they have been among older users. Moreover, this study adds to the prior findings that time affects user perceptions of interface aesthetics (Lee and Koubek, 2011; Hartmann *et al.*, 2008) in such a way that users become more selective over time.

In terms of the adaptive decision-making theory, which has been advanced only minimally in the context of aesthetics in interaction design, users have been found to apply a tradeoff strategy by weighting different attributes of designs to an extent by the users' background (Hartmann *et al.*, 2007a). Drawing from this theoretical framework, as both younger women and men nowadays may have received greater exposure to technology compared with the older generation, the background of users seems to have become more homogenous, thus leading to similar perceptions of visual aesthetics according to the theory. Thus, the decision-making strategies seem to have unified as well, which may imply a change to the psychological patterns of making decisions altogether.

Previous studies that have employed adaptive decision-making theory in similar contexts (Hartmann *et al.*, 2007a, b, 2008) posit that aesthetics should be matched to user profile. We contribute to the literature by offering deeper insight on how the cultural atmosphere seems to be changing user preferences and decision-making behavior to predict their intention towards the judgment of aesthetic appeal, thus aiding scholars to revisit theories on decision-making and aesthetic appeal. Particularly, as the adaptive decision-making theory has not been applied widely and recently in similar studies, a critical approach could be adapted in order to systematically build the theory for the development of new hypotheses. In conclusion, we have demonstrated that aesthetics is a component of design quality that is susceptible to the user's decision-making strategies. Implementing this theoretical framework shows evidence that user perception is a complex construct that requires the understanding of deeper behavioral meaning.

5.2 Practical implications

This study adds to the existing literature of designing graphical user interface elements relating to demographic effects. Aesthetic appeal is a complex matter, nonetheless, some practical implications can be made on the basis of the findings.

Design implication 1: The results suggest that younger users in general, as well as older women, tend to be more critical towards icon aesthetics. Thus, in order to visually appeal to the tastes of younger audiences and women, focusing on creating high quality designs (i.e. *high graphical fidelity*) is recommended, as the hedonic aspects need to be catered to across these demographic factors.

Design implication 2: Expectedly, time affects perceptions in that novice users perceive icons as more excellent than experienced users. Therefore, in order to visually appeal to more experienced users, designers may have to put in more effort and creativity.

Design implication 3: Overall, gender differences among younger users seem to be minimizing and therefore gender-neutral options could be considered in future design processes. However, the perceptions of icons change especially for younger women in that icons are seen as more concrete and exciting over time. Hence, practitioners could benefit from integrating young female users to interfaces at an early stage to increase the aforementioned effects.

The results suggest that user perceptions are subjective and thus age, gender and time have relatively little effect on how users evaluate icon aesthetics. However, these implications inform what kinds of aesthetic perceptions graphical user interface elements (i.e. *icons*) should be brought to evoke. This knowledge can then be adapted in establishing segmentation models for the design of adapted user interfaces.

5.3 Limitations and future research

This study was one of the first attempts to understand how demographic factors affect user perceptions of GUI element aesthetics by utilizing game app icons as data collection material. However, there are some limitations that should be acknowledged.

Game app icons were used as study material to maximize internal validity. This poses 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.

The data was gathered via an online survey that was advertised on Finnish student organizations' mailing lists, thus the sample can be considered fairly homogenous. The majority of the respondents are from the same age group and come from a similar cultural background, which could affect perceptions in the study. Moreover, the sample in this study is a nonprobability convenience sample, therefore it is not necessarily representative of all app store users. In future research, a more diverse sample should be gathered in order to gain perspective on factors related to age and cultural differences.

As is commonplace within the industry, actual data on app store usage was not available, thus the measurement used in this study reports intended behavior with a vignette style experiment setting. This may have an impact on the generalizability of the findings. Moreover, as the results consist of perceptions measured by quantitative means, the findings may be considered ambiguous with underlying biases. Therefore, a qualitative approach would be beneficial in order to gain a deeper understanding of the topic in further studies. Additionally, an even more authentic experiment setting could be composed.

6. Conclusion

This study replicated prior literature in the sense that paying close attention to visual aesthetics is important, especially when targeting experienced users, young audiences and women. Knowledge about demographic effects relating to how GUI elements (i.e. *app icons*) are perceived is scarce, therefore, insight into the topic is valuable for deciding on effective design processes. Considering the changing cultural atmosphere, especially relating to gender and age in the domain of technology, insight into the topic is valuable. The current undertaking shows that technology adoption advances at a tremendous pace, which blurs the boundaries of aesthetics between people despite their age, gender and habits in daily life.

References

- Ahmed, S.U., Al Mahmud, A. and Bergaust, K. (2009), "Aesthetics in human-computer interaction: views and reviews", in Jacko, J.A. (Ed.), *Human-Computer Interaction, New Trends, HCI 2009, Lecture Notes in Computer Science*, Springer, Berlin, Vol. 5610.
- Aiken, L.S. and West, S.G. (1991), *Multiple Regression: Testing and Interpreting Interactions*, Sage, London.
- Arend, U., Muthig, K.P. and Wandmacher, J. (1987), "Evidence for global superiority in menu selection by icons", *Behaviour and Information Technology*, Vol. 6 No. 4, pp. 411-426, doi: 10.1080/01449298708901853.
- Beresford, B. and Sloper, T. (2008), *Understanding the Dynamics of Decision-Making and Choice: A Scoping Study of Key Psychological Theories to Inform the Design and Analysis of the Panel Study*, Social Policy Research Unit, University of York, York.
- Blankenberger, S. and Hahn, K. (1991), "Effects of icon design on human-computer interaction", *International Journal of Man-Machine Studies*, Vol. 35 No. 3, pp. 363-377, doi: 10.1016/S0020-7373(05)80133-6.
- Boiano, S., Borda, A., Bowen, J., Faulkner, X., Gaia, G. and Mcdaid, S. (2006), "Gender issues in HCI design for web access", in Kurniawan, S. and Zaphiris, P. (Eds), *Advances in Universal Web Design and Evaluation: Research, Trends and Opportunities*, IGI Global, London, pp. 116-153.

-
- Chartboost (2015), "Power-up report–July 2015", available at: <https://chartboost.s3.amazonaws.com/blog/power-up-report-july-2015-building-an-empire-mobile-strategy-games.pdf> (accessed 14 September 2018).
- Choi, J.H. and Lee, H.J. (2012), "Facets of simplicity for the smartphone interface: a structural model", *International Journal of Human-Computer Studies*, Vol. 70 No. 2, pp. 129-142, doi: 10.1016/j.ijhcs.2011.09.002.
- Chung, J.E., Park, N., Wang, H., Fulk, J. and McLaughlin, M. (2010), "Age differences in perceptions of online community participation among non-users: an extension of the technology acceptance model", *Computers in Human Behavior*, Vol. 26 No. 6, pp. 1674-1684.
- Creusen, M.E.H. and Schoormans, J.P.L. (2005), "The different roles of product appearance in consumer choice", *Journal of Product Innovation Management*, Vol. 22 No. 1, pp. 63-81, doi: 10.1111/j.0737-6782.2005.00103.x.
- Creusen, M.E.H., Veryzer, R.W. and Schoormans, J.P.L. (2010), "Product value importance and consumer preference for visual complexity and symmetry", *European Journal of Marketing*, Vol. 44 Nos 9/10, pp. 1437-1452, doi: 10.1108/03090561011062916.
- Creusen, M.E.H. (2010), "The importance of product aspects in choice: the influence of demographic characteristics", *Journal of Consumer Marketing*, Vol. 27 No. 1, pp. 26-34, doi: 10.1108/07363761011012921.
- Cyr, D. (2009), "Gender and website design across cultures", *Paper Presented at the 17th European Conference on Information Systems*, Verona, Italy, pp. 279-291.
- Debevc, M., Meyer, B., Donlagic, D. and Svecko, R. (1996), "Design and evaluation of an adaptive icon toolbar", *User Modeling and User-Adapted Interaction*, Vol. 6 No. 1, pp. 1-21, doi: 10.1007/BF00126652.
- Dewar, R. (1999), "Design and evaluation of public information symbols", in Zwaga, H.J.G., Boersema, T. and Hoonhout, H.C.M. (Eds), *Visual Information for Everyday Use*, Taylor & Francis, London, pp. 285-303.
- Gait, J. (1985), "An aspect of aesthetics in human-computer communications: pretty windows", *IEEE Transactions on Software Engineering*, Vol. 11 No. 8, pp. 714-717, doi: 10.1109/TSE.1985.232520.
- Genuine (2013), "Gender-inclusive user interface guidelines", available at: <http://genuine.ict.tuwien.ac.at/unterlagen/Guidelines.pdf> (accessed 27 May 2019).
- Goonetilleke, R.S., Shih, H.M., On, H.K. and Fritsch, J. (2001), "Effects of training and representational characteristics in icon design", *International Journal of Human-Computer Studies*, Vol. 55 No. 5, pp. 741-760, doi: 10.1006/ijhc.2001.0501.
- Hartmann, J., Sutcliffe, A. and De Angeli, A. (2007a), "Towards a theory of user judgment of aesthetics and user interface quality", *ACM Transactions on Computer-Human Interaction*, Vol. 15 No. 4, doi: 10.1145/1460355.1460357.
- Hartmann, J., Sutcliffe, A. and De Angeli, A. (2007b), "Investigating attractiveness in web user interfaces", *Paper Presented at the 25th Annual SIGCHI Conference on Human Factors in Computing Systems*, San Jose, USA, pp. 387-396.
- Hartmann, J., De Angeli, A. and Sutcliffe, A. (2008), "Framing the user experience: information biases on website quality judgement", *Paper Presented at the 26th Annual SIGCHI Conference on Human Factors in Computing Systems*, Florence, Italy, pp. 855-864.
- Hassenzahl, M., Burmester, M. and Koller, F. (2003), "AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität [AttracDiff: a questionnaire to measure perceived hedonic and pragmatic quality]", in Ziegler, J. and Szwillus, G. (Eds), *Mensch and Computer 2003–Interaktion in Bewegung*, B. G. Teubner, Stuttgart, pp. 187-196.
- Henry, P. (2002), "Systematic variation in purchase orientations across social classes", *Journal of Consumer Marketing*, Vol. 19 No. 5, pp. 424-438, doi: 10.1108/07363760210437641.
- Hoffman, L.W. (1972), "Early childhood experiences and women's achievement motives", *Journal of Social Issues*, Vol. 28 No. 2, pp. 129-155.

-
- Hou, K.-C. and Ho, C.-H. (2013), "A preliminary study on aesthetic of apps icon design", *Paper Presented at the 5th International Congress of International Association of Societies of Design Research*, Tokyo, Japan, pp. 3845-2856.
- Hsiao, K.-L. and Chen, C.-C. (2016), "What drives in-app purchase intention for mobile games? An examination of perceived values and loyalty", *Electronic Commerce Research and Applications*, Vol. 16, pp. 18-29, doi: 10.1016/j.elerap.2016.01.001.
- Hsieh, M.-H., Pan, S.-L. and Setiono, R. (2004), "Product-, corporate-, and country-image dimensions and purchase behavior: a multicountry analysis", *Journal of the Academy of Marketing Science*, Vol. 32 No. 3, pp. 251-270, doi: 10.1177/0092070304264262.
- Huang, S.M., Shieh, K.K. and Chi, C.F. (2002), "Factors affecting the design of computer icons", *International Journal of Industrial Ergonomics*, Vol. 29 No. 4, pp. 211-218, doi: 10.1016/S0169-8141(01)00064-6.
- Huang, S. (2013), "Usability and GUI design and principles", available at: https://www.uio.no/studier/emner/matnat/ifi/INF5120/v13/undervisningsmateriale/f04-2013_0402-3-designguidelines_additional_info.pdf (accessed 27 May 2019).
- Isherwood, S.J., McDougall, S.J.P. and Curry, M.B. (2007), "Icon identification in context: the changing role of icon characteristics with user experience", *Human Factors*, Vol. 49 No. 3, pp. 465-476, doi: 10.1518/001872007X200102.
- Jennings, M. (2000), "Theory and models for creating engaging and immersive ecommerce websites", *Paper Presented at the 2000 ACM SIGCPR Conference on Computer Personnel Research*, New York, USA, pp. 77-85, doi: 10.1145/333334.333358.
- Johnson, J. and Finn, K. (2017), *Designing User Interfaces for an Aging Population: Towards Universal Design*, Morgan Kaufmann, Amsterdam.
- Jylhä, H. and Hamari, J. (2019), "An icon that everyone wants to click: how perceived aesthetic qualities predict app icon successfulness", *International Journal of Human-Computer Studies*, Vol. 130, pp. 73-85, doi: 10.1016/j.ijhcs.2019.04.004.
- Jylhä, H. and Hamari, J. (2020), "Development of measurement instrument for visual qualities of graphical user interface elements (VISQUAL): a test in the context of mobile game icons", *User Modeling and User-Adapted Interaction*, Vol. 30 No. 5, pp. 949-982, doi: 10.1007/s11257-020-09263-7.
- KnowItAll Ninja (2016), "Exploring user interface design principles and project planning techniques", available at: <https://www.pearsonschoolsandfecolleges.co.uk/AssetsLibrary/SECTORS/FurtherEducationColleges/SUBJECT/dit/dit-student-book/btec-techaward-dit-sb-9781292208374.pdf> (accessed 27 May 2019).
- Kurosu, M. and Kashimura, K. (1995), "Apparent usability vs. inherent usability", *Paper Presented at the CHI 95 Conference Companion on Human Factors in Computing Systems*, New York, USA, pp. 292-293, doi: 10.1145/223355.223680.
- Lavie, T. and Tractinsky, N. (2004), "Assessing dimensions of perceived visual aesthetics of web sites", *International Journal of Human-Computer Studies*, Vol. 60 No. 3, pp. 269-298, doi: 10.1016/j.ijhcs.2003.09.002.
- Lee, S. and Koubek, R.J. (2011), "The impact of cognitive style on user preference based on usability and aesthetics for computer-based systems", *International Journal of Human-Computer Studies*, Vol. 27 No. 11, pp. 1083-1114, doi: 10.1080/10447318.2011.555320.
- Leong, L.-Y., Ooi, K.-B., Chong, A.Y.-L. and Lin, B. (2013), "Modeling the stimulators of the behavioural intention to use mobile entertainment: does gender really matter?", *Computers in Human Behavior*, Vol. 29 No. 5, pp. 2109-2121, doi: 10.1016/j.chb.2013.04.004.
- Lin, C.L. and Yeh, J.T. (2010), "Marketing aesthetics on the web: personal attributes and visual communication effects", *Paper Presented at the 5th IEEE International Conference on Management of Innovation and Technology*, IEEE, pp. 1083-1088.

-
- Linux Information Project (2004), "GUI definition", available at: <http://www.linfo.org/gui.html> (accessed 28 May 2019).
- Liu, D. and Guo, X. (2017), "Exploring gender differences in acceptance of mobile computing devices among college students", *Information Systems and E-Business Management*, Vol. 15 No. 1, pp. 197-223, doi: 10.1007/s10257-016-0315-x.
- Maity, R., Uttav, A., Gourav, V. and Bhattacharya, S. (2015), "A non-linear regression model to predict aesthetic ratings of on-screen images", *Paper Presented at the Annual Meeting of the Australian Special Interest Group for Computer Human Interaction*, OZCHI 2015, Parkville, pp. 44-52, doi: 10.1145/2838739.2838743.
- McDougall, S.J.P. and Reppa, I. (2008), "Why do I like it? The relationships between icon characteristics, user performance and aesthetic appeal", *Paper Presented at the Human Factors and Ergonomics Society 52nd Annual Meeting*, New York, USA, pp. 1257-1261, doi: 10.1177/154193120805201822.
- McDougall, S.J.P. and Reppa, I. (2013), "Ease of icon processing can predict icon appeal", in *Proceedings of the 15th International Conference on Human-Computer Interaction*, Las Vegas, pp. 575-584, doi: 10.1007/978-3-642-39232-0_62.
- McDougall, S.J.P., Curry, M.B. and de Bruijin, O. (1999), "Measuring symbol and icon characteristics: norms for concreteness, complexity, meaningfulness, familiarity, and semantic distance for 239 symbols", *Behavior Research Methods, Instruments, and Computers*, Vol. 31 No. 3, pp. 487-519, doi: 10.3758/BF03200730.
- McDougall, S.J.P., de Bruijin, O. and Curry, M.B. (2000), "Exploring the effects of icon characteristics on user performance: the role of icon concreteness, complexity, and distinctiveness", *Journal of Experimental Psychology: Applied*, Vol. 6 No. 4, pp. 291-306, doi: 10.1037/1076-898X.6.4.291.
- McDougall, S.J.P., Reppa, I., Kulik, J. and Taylor, A. (2016), "What makes icons appealing? The role of processing fluency in predicting icon appeal in different task contexts", *Applied Ergonomics*, Vol. 55, pp. 156-172, doi: 10.1016/j.apergo.2016.02.006.
- Miniukovich, A. and De Angeli, A. (2014), "Visual impressions of mobile app interfaces", *Paper Presented at the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, Helsinki, Finland, pp. 31-40, doi: 10.1145/2639189.2641219.
- Minton, H.L. and Schneider, F.W. (1980), *Differential Psychology*, Waveland Press, Prospect Heights, IL.
- Morris, M.G. and Venkatesh, V. (2000), "Age differences in technology adoption decisions: implications for a changing work force", *Personnel Psychology*, Vol. 53 No. 2, pp. 375-403.
- Morris, M.G., Venkatesh, V. and Ackerman, P.L. (2005), "Gender and age differences in employee decisions about new technology: an extension to the theory of planned behavior", *IEEE Transactions on Engineering Management*, Vol. 52 No. 1, pp. 69-84.
- Moyes, J. and Jordan, P.W. (1993), "Icon design and its effect on guessability, learnability, and experienced user performance", in Alty, J.D., Diaper, D. and Gust, S. (Eds), *People and Computers VIII*, Cambridge University Society, Cambridge, pp. 49-59.
- Ngo, D.C.L., Samsudin, A. and Abdullah, R. (2000), "Aesthetic measures for assessing graphic screens", *Journal of Information Science and Engineering*, Vol. 16 No. 1, pp. 97-116.
- Ngo, D.C.L., Teo, L.S. and Byrne, J.G. (2003), "Modelling interface aesthetics", *Information Sciences*, Vol. 152 No. 1, pp. 25-46, doi: 10.1016/S0020-0255(02)00404-8.
- Ngo, D.C.L. (2001), "Measuring the aesthetic elements of screen designs", *Displays*, Vol. 22 No. 3, pp. 73-78, doi: 10.1016/S0141-9382(01)00053-1.
- Norman, D.A. (2004), *Emotional Design: Why We Love (Or Hate) Everyday Things*, Basic Books, New York.
- Overby, E. and Sabyasachi, M. (2014), "Physical and electronic wholesale markets: an empirical analysis of product sorting and market function", *Journal of Management Information Systems*, Vol. 31 No. 2, pp. 11-46, doi: 10.2753/MIS0742-1222310202.

- Oyibo, K. and Vassileva, J. (2017), "The interplay of aesthetics, usability and credibility in mobile website design and the moderation effect of gender", *Journal on Interactive Systems*, Vol. 8 No. 2, pp. 4-19.
- Oyibo, K., Ali, Y.S. and Vassileva, J. (2016), "An empirical analysis of the perception of mobile website interfaces and the influence of culture", in Orji, R., Reisinger, M., Busch, M., Dijkstra, A., Stibe, A. and Tscheligi, M. (Eds), *Proceedings of the Personalization in Persuasive Technology Workshop, Persuasive Technology 2016*, Salzburg, Austria, pp. 44-56.
- Oyibo, K., Adaji, I. and Vassileva, J. (2018), "The effect of age and information design on the perception of visual aesthetic", *Paper Presented at the British Human Computer Interaction Workshop*, Belfast, UK. doi: 10.14236/ewic/HCI2018.208.
- Pappas, I.O., Mikalef, P., Giannakos, M.N. and Kourouthanassis, P.E. (2019), "Explaining user experience in mobile gaming applications: an fsQCA approach", *Internet Research*, Vol. 29 No. 2, pp. 293-314, doi: 10.1108/IntR-12-2017-0479.
- Payne, J., Bettman, J. and Johnson, E. (1993), *The Adaptive Decision Maker*, Cambridge University Press, Cambridge. doi: 10.1017/CBO9781139173933.
- Reinecke, K. and Gajos, K.Z. (2014), "Quantifying visual preferences around the world", *Paper Presented at the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*, ACM, pp. 11-20.
- Rogers, Y. and Osborne, D.J. (1987), "Pictorial communication of abstract verbs in relation to human-computer interaction", *British Journal of Psychology*, Vol. 78 No. 1, pp. 99-112, doi: 10.1111/j.2044-8295.1987.tb02229.x.
- Rousseau, G.K., Lamson, N. and Rogers, W.A. (1998), "Designing warnings to compensate for age-related changes in perceptual and cognitive abilities", *Psychology and Marketing*, Vol. 15 No. 7, pp. 643-662.
- Salimun, C., Purchase, H.C., Simmons, D. and Brewster, S. (2010), "The effect of aesthetically pleasing composition on visual search performance", *Paper Presented at the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, Reykjavik, Iceland, ACM, pp. 422-431, doi: 10.1145/1868914.1868963.
- Salman, Y.B., Kim, Y. and Cheng, H. (2010), "Senior-friendly icon design for the mobile phone", *Paper Presented at the 6th International Conference on Digital Content, Multimedia Technology and its Applications (IDC 2010)*, IEEE, Seoul, pp. 103-108.
- Sarsam, S.M. and Al-Samarraie, H. (2018), "Towards incorporating personality into the design of an interface: a method for facilitating users' interaction with the display", *User Modeling and User-Adapted Interaction*, Vol. 28 No. 1, pp. 75-96, doi: 10.1007/s11257-018-9201-1.
- Shaikh, A.D. (2009), "Know your typefaces! Semantic differential presentation of 40 onscreen typefaces", *Usability News*, Vol. 11, pp. 23-65.
- Shu, W. and Lin, C.-S. (2014), "Icon design and game app adoption", in *Proceedings of the 20th Americas Conference on Information Systems*, Georgia.
- Smith, G.E. (1995), "Framing product design: using design communication to facilitate user learning", *Journal of Business and Industrial Marketing*, Vol. 10 No. 5, pp. 6-21, doi: 10.1108/08858629510103860.
- Sun, H. and Zhang, P. (2006), "The role of moderating factors in user technology acceptance", *International Journal of Human-Computer Studies*, Vol. 64 No. 2, pp. 53-78.
- Sweet, S.A. and Grace-Martin, K.A. (2011), *Data Analysis with SPSS: A First Course in Applied Statistics*, 4th ed., Pearson, London.
- Thüring, M. and Mahlke, S. (2007), "Usability, aesthetics and emotions in human-technology interaction", *International Journal of Psychology*, Vol. 42 No. 4, pp. 253-264, doi: 10.1080/00207590701396674.
- Tractinsky, N., Katz, A.S. and Ikar, D. (2000), "What is beautiful is useable", *Interacting with Computers*, Vol. 13 No. 2, pp. 127-145, doi: 10.1016/S0953-5438(00)00031-X.

- Tuch, A.N., Bargas-Avila, J.A. and Opwis, K. (2010), "Symmetry and aesthetics in website design: it's a man's business", *Computers in Human Behavior*, Vol. 26 No. 6, pp. 1831-1837, doi: 10.1016/j.chb.2010.07.016.
- Venkatesh, V. and Morris, M.G. (2000), "Why don't men ever stop to ask for directions? Gender, social influence, and their role in technology acceptance and usage behavior", *MIS Quarterly*, Vol. 24 No. 1, pp. 115-139.
- Venkatesh, V., Morris, M.G. and Ackerman, P.L. (2000), "A longitudinal field investigation of gender differences in individual technology adoption decision-making processes", *Organizational Behavior and Human Decision Processes*, Vol. 83 No. 1, pp. 33-60.
- Wallendorf, M. and Arnould, E.J. (1988), "My favorite things: a cross-cultural inquiry into object attachment, possessiveness, and social linkage", *Journal of Consumer Research*, Vol. 14 No. 4, pp. 531-547.
- Zen, M. and Vanderdonckt, J. (2016), "Assessing user interface aesthetics based on the inter-subjectivity of judgment", *Paper Presented at the 30th International BCS Human Computer Interaction Conference*, BCS, Swindon. doi: 10.14236/ewic/HCI2016.25.

Corresponding author

Henrietta Jylhä can be contacted at: henrietta.jylha@gmail.com

