THE DESIGN JOURNAL

VOLUME 0, ISSUE 0
PP 1–25

REPRINTS AVAILABLE
DIRECTLY FROM
THE PUBLISHERS

PHOTOCOPYING
PERMITTED BY
LICENSE ONLY

What Makes Good Design? Revealing the Predictive Power of Emotions and Design Dimensions in Non-Expert Design Vocabulary

Chaehan So

International Design School for Advanced Studies, Hongik University, Seoul, South Korea

ABSTRACT This paper investigates how non-experts perceive digital design, and which psychological dimensions are underlying this perception of design. It thus constructs a measurement instrument to analyse user response to online displayed design and to predict design preference. Study 1 let non-

The Design Journal DOI: 10.1080/14606925.2019.1589204

Routledge
Taylor & Francis Group

experts rank the usefulness of 115 adjectives for describing good design in an online survey (n = 305). This item pool was condensed to 12 design descriptive and five emotion items. Exploratory factor analysis revealed the four underlying psychological dimensions Novelty, Energy, Simplicity and Tool. Study 2 (n = 1955) tested Study 2's model in three real-world design projects. Emotions clearly outperformed the best design descriptive dimensions (Novelty and Tool) in predicting users' design preference (Net Promoter Score) with $\beta = .82$. Study 3 (n = 1955) confirmed Study 2's results by several machine learning algorithms (neural networks, gradient boosting machines, random forests) with cross-validation. This measurement instrument benefits designers to implement a participatory design thinking process with users.

KEYWORDS: design psychology, design thinking, user-centred design, participatory design, machine learning

Introduction

What is good design? Because of its simplicity, research has extensively and endlessly tackled this question from many different angles. Although a complete coverage of attempted answers appears impossible, some can be considered as established main ideas and therefore mentionable. From fine arts and the philosophical stream of idealism, the question on the essence of beauty established the concept of aesthetics (Karvonen 2000; Ann and Stankiewicz 1987). In juxtaposition of this view, design and engineering disciplines put the aspects of usefulness and practicality into focus (Tuch et al. 2012). These pragmatic aspects had been analysed before in the disciplines ergonomics and human factors with physical products and machines (Wickens, Gordon, and Liu 2004; Sharit 2006). With the advent of the internet, a new research stream emerged that primarily focused on the usability of digital products (e.g. websites, mobile apps, IoT interfaces). This research stream expanded from the initial focus of usability in diverse directions and bred several disciplines that were semantically largely overlapping despite different names: user interaction design (Ju and Leifer 2008; Biskjaer and Halskov 2014), user experience (UX) design (Marsh, Wong, and Carriazo 2005; Hassenzahl and Tractinsky 2006), and human-computer interaction (HCI) (Rosson and Carroll 2002; Jacko 2012), which became an umbrella term for the preceding terms. In the late 2000s, the newborn (or new-branded) methodology design thinking re-emphasized the importance of user needs and human needs in the approaches user-centred design (Norman and Draper 1986; Abras, Maloney-Krichmar, and Preece 2004) and human-centred design (Büttner and Röcker 2016; Hanington 2003).

Many integrative approaches emerged that analysed the interplay between aesthetic and usability aspects (Hassenzahl 2004; Tuch et al. 2012; Sonderegger and Sauer 2010) or the crossover of psychology (Hollan, Hutchins, and Kirsh 2000; Card 2017), in particular for understanding the role of emotions in interaction design (Desmet, Overbeeke, and Tax 2001; Desmet 2012).

The continuous advancement in these disciplines sometimes makes us forget that they require substantial expert knowledge for evaluating design. Experts distinguish themselves from non-experts by domain-specific knowledge that frames their perception (Chi 2006, 168–170). For example, experts describe the behaviour of objects more by higher-order laws whereas non-experts describe this behaviour more in concrete terms (Gredler 1997), they evaluate risks differently than non-experts (Bostrom 1997), and they apply specific cognitive heuristics (Yilmaz and Seifert 2009, 2011). Experts sometimes use their knowledge to undermine non-experts (Cook, Pieri, and Robbins 2004; Turner 2001). In many domains, however, the descriptions by experts are not more accurate than by non-experts, e.g. for annotations in natural language tasks (Snow et al. 2008), evaluations of music hall acoustics (Galiana, Llinares, and Page 2012), or valuations of material goods (Sakalaki and Kazi 2007).

Many discipline-specific methods and measurement instruments have been developed and used for design evaluation purposes (Chi 2006; Bangor, Kortum, and Miller 2008, 2009; Hassenzahl, Burmester, and Koller 2003). Hence, the disparate phenomenology of design has been extensively covered with analysis tools created by domain experts so far.

Change of perspective

The progress in the past decades of specialization, however, has left the possibility of a generalized view under-investigated. What might have been missed out is that because non-experts have different perceptions of design (Hekkert and Van Wieringen 1996), they also have different design vocabulary than the design experts. This non-expert vocabulary is bound to scientific investigation no less than any other design discipline.

Therefore, the current work proposes to leave the established expert-driven, discipline-oriented realm of research behind and to approach the central question of good design from the perspective of non-experts. Following this route requires a domain-free, eclectic, and user-centric approach. This requirement leads to rigorously applying psychometrics (Kline 1998), the science of psychological measurement, because it has established a long tradition of systematically reducing biases from various sources (R. J. Cohen, Swerdlik, and Phillips 1996).

Abandoning the domain expert realm in favour of understanding common people's perspectives with no apparent expert education has been labelled as research ‘in the wild’. This emerged in the late 2000s as a popular new term in HCI research (Minh et al. 2011; Chamberlain et al. 2012; Li et al. 2016) but has been known long before in psychology (e.g. Hutchins 1995). In recent years, artificial intelligence has produced many studies (Huang et al. 2007; Zhu and Ramanan 2012) using data in the wild, i.e. people’s faces or poses from internet-based sources like the most popular social media portals (e.g. Facebook, Twitter, Instagram). Research in the wild aims to acquire real-world data to enable the practical applicability of developed technical frameworks and representative conclusions from data analysis.

In a symbiotic manner, sample acquisition in the wild goes along with recruiting non-experts in the design field to understand a purely user-centred perspective. Such analysis of non-expert views is often referred to as ‘lay perspective’ in psychological and consumer research, and non-experts are referred to as ‘lay people’. The views of lay people often render insights on the general intuitive understanding of people towards a relevant topic. This research has been conducted across a wide range of aspects including lay people’s perceptions of facial aesthetics (Flores-Mir et al. 2004), risk (Bostrom 1997), ethnic prejudice (Hodson and Esses 2005), and lay beliefs on advertising (Friestad and Wright 1995) and on satisfaction and performance (Fisher 2003).

The present work determines its research context on the assumption that the most frequent scenario in the wild in which people encounter design works is online – from reading blogs, frequenting social media portals, or buying products online. In these media, they see visualized digital 2D design work in a random and unintentional manner, hence relatively unbiased from preconceived perceptions caused, e.g., by brand loyalty. This scenario prevents biases that are present in situations when people are interviewed due to multiple biases between interviewer and interviewee.

Research questions

Understanding non-expert design vocabulary methodology is a new route taken in design research. It aims to replace an expert-driven debate on design values by the direct insight into the perception of those design aspects that explain non-expert design preference. Design preference in the present work shall be defined as the degree to which people recommend a design work to others as an example of good design.

The preceding considerations lead to the following research questions:

RQ1: How do non-experts perceive design?

RQ2: How well does non-expert design vocabulary predict design preference?

To answer RQ1, the present work elaborates in Study 1 the non-experts' intuitive understanding of visual design by an extensive study of their vocabulary usage, condenses it to a minimal set and derives the psychological components underlying the non-expert perception. Thus, Study 1's outcome is a new measurement instrument for capturing non-expert design perception which Study 2 validates in real-world design projects.

Furthermore, Study 2 tests how well this new measurement instrument can predict design preference with linear models, thus providing a partial answer to RQ2.

To fully answer RQ2, Study 3 compares the prediction performance of Study 2's linear model with a selection of modern data science models from artificial intelligence. Study 3 thus aims to maximize the measurement instrument's usefulness for predicting design preference by finding the best predictive model.

Building an integrative picture by methodology

The following explains why the methodologies in the present work were chosen, and how they integrate into a cohesive rationale to answer the research questions.

The methodological frameworks for Study 1 and 2 are drawn from the field of psychology because 1) investigating a phenomenon of perception is a main subject of psychological science (Zimbardo, Johnson, and McCann 2012) and 2) understanding an underlying psychological pattern is intrinsic to psychological methodology (for excellent overviews cf. Tabachnick and Fidell, 2006; Giles 2002).

Quantitative psychological methodology allows to analyse perceptive patterns by exploratory factor analysis and to verify psychological models by hypothesis testing – together, these methods allow to make justified generalizations on the ground of commonly agreed statistical techniques (Tabachnick and Fidell 2006, 17–27).

The role of qualitative research methodology – contrary to common misconception – is not of a competing but rather of a complementary one to quantitative methodology. Therefore, Study 1's construction of a quantitative measurement instrument is preceded by a qualitative effort of collecting input for potential survey questions called items.

After item collection, Study 1's scale construction aims to analyse participants' item ratings to create a reduced set of items. The applied criterion is the so-called factor structure, i.e. a pattern of item sets that are assumed to correspond to underlying psychological dimensions (Giles 2002, 121–129).

The statistical tests in Study 2 represent the most common form of hypothesis testing in psychology which assumes a linear model underlying the observations. This analysis can be performed both on the individual item level and on the aggregated component level. The hypothesis tests on a linear model thus translate into the question whether people’s design preference increases with higher scores on a particular item or component.

The measure for design preference was derived from the NPS, the Net Promoter Score (Reichheld 2003). The NPS measures the recommend intention of a consumed product. Subsequent research rejected Reichheld's (2003) claim for the NPS to be the ‘single most reliable indicator’ for firm growth and customer loyalty by showing that customer satisfaction performed similarly well (Keiningham et al. 2008, 86; Keiningham, Aksoy, and Cooil 2008). Despite this criticism, the NPS is still one of the most widely used measures of customer satisfaction to date.

The conducted hypothesis tests should be validated in new samples of future studies. Yet, cross-validation of the same sample is also possible if we extend our methodological toolbox to a data science discipline known as machine learning, a subset of artificial intelligence. In k-fold cross-validation, machine learning models reserve an iterative portion of the training data to validate the prediction estimate (Ghatak 2017). As final validation step, the model’s prediction performance is measured on a hold-out subset of the original data, the testing set. As this data was not used during the model training, the testing set performance measurement is unbiased from the training data.

Furthermore, machine learning as a discipline provides a large pool of predictive models, i.e. models that predict a categorical variable in a classification model or a numerical variable in a regression model (Kuhn and Johnson 2013). Although the underlying models differ widely, they can be abstracted as a black box and thus be compared by the same metric – accuracy for classification or root mean squared error (RMSE) for regression models.

Taken together, the value of applying several machine learning models is two-fold because they show whether they can predict design preference better than linear regression, and whether their prediction performance differs on unseen data.

Study 1

The goal of Study 1 was to elaborate a wide non-expert vocabulary of design-descriptive adjectives which non-experts use intuitively. Based on a pool of 115 items from qualitative and quantitative research, it builds a measurement instrument for design perception to answer RQ1 (How do non-experts perceive design?).

Method

Participants

Study 1 acquired a sample of 305 participants (137 male, 168 female, mean age = 33.2 years, SD = 14.2 years) for a web-based questionnaire. The sample’s ethnicity consisted of 164 Caucasians, 101 Asians, 22 Latinos/Hispanics, 14 African Americans/Blacks, and

4 multi-racial. The sample's occupation structure was dominated by non-design professionals (108) and non-design students (57), followed by a wide variety of non-design-related occupations including six musicians.

Item pool generation

Study 1 generated an item pool of 115 descriptive adjectives that addressed a broad range of design aspects and emotional experience. The first part thereof, 28 items (24%), was extracted from the following scales drawn from psychological and design research.

Usability. Extensive research by Hassenzahl and colleagues (Hassenzahl 2001; Hassenzahl, Burmester, and Koller 2003) provided a validated source of usability items. Several items were removed due to reduced linguistic familiarity or lack of generalizability outside the usability context, and only the positive element of the semantic differential was used (Hassenzahl, Burmester, and Koller 2003, 192, Table 1). The following 12 items were selected – HQ-S subscale: exciting, creative, original, innovative, courageous; HQ-I subscale: connecting, integrating, precious, stylish, professional; EQ subscale: simple, clear.

Emotions. The emotion items were derived from three sources. First, positive psychology research investigates positive emotions, i.e. emotions that promote flourishing over time, and predict subjective well-being. Three of the positive emotions found by Fredrickson (2001) were selected – happy, proud, loving. Second, consumer research investigates consumption emotions, i.e. emotions that are related to consumer satisfaction and evoked during and after consumption of products, e.g. interest, joy, surprise, sadness and anger (Westbrook and Oliver, 1991, 87, Table 1). Subsequent research by Mano and Oliver (1993) revealed that post-consumption product evaluation can be categorized into a utilitarian and hedonic judgement that causes affect and product satisfaction (p.455, Figure 2). From this consumer research, the following five emotions were selected: excited, inspired, aroused, pleased, satisfied. Third, design research investigates emotions elicited by design works mostly in the product or industrial design context. Desmet identified the most frequently experienced positive emotions with products (P. M. A. Desmet 2012, 8, Table 3). From this design research, seven items were selected: amused, relaxed, confident, desire, energized, pleasantly surprised, anticipating. In addition, the item interested was included because it revealed to be the most relevant emotion in user experience due to high frequency, positive relationship to well-being, and motivating role for learning (Yoon, Desmet, and van der Helm 2012).

In summary, a total of 16 emotion-related items was gathered by the above procedure, namely: happy, proud, loving, excited, inspired, aroused, pleased, satisfied, amused, relaxed, confident, desire, energized, pleasantly surprised, anticipating, interested.

Expert interviews

The second part of the item pool, 87 items (76%), was created by qualitative methodology from interviews with eight experts (five design professors, two design thinking consultants, one musician with design experience). Among the expert items, multiple mentions were removed. Then, lemmatization was applied to sort out derivatives of the same word (e.g. inspiring, inspirational). From the remainder, three experts extracted 20 items (26%) as the final selection by the criteria familiarity and usage frequency in common language reflecting non-expert vocabulary. In case of doubt, the experts applied linguistic familiarity as a selection criterion according to the guideline: if two items were semantically similar, the item was chosen by a higher number of google search hits, corresponding to higher familiarity with this word.

Survey

The final survey of Study 1 went online for six weeks between 1 February and 15 March 2017. Participants were instructed to answer the question: 'How well do the following adjectives describe the quality of design work? In other words, does it make sense to say "this design is … " to express it is good design?'. This question was evaluated for each design descriptive and emotion item on a 5-point Likert scale (not at all – extremely well).

To ensure sufficient sample acquisition, the study applied a parallel launch of several online recruitment strategies: the survey was published in the Facebook, LinkedIn, Xing, Twitter, 'Call for Participants', and 'Psychological Research on the Net'. In addition, a Google AdWords campaign was launched on 4 March 2017 that acquired 91,331 impressions and 2,212 clicks (2.42%) during five weeks.

Dimensionality reduction

To reduce the item pool, items showing positive skewness were eliminated because this meant that the majority of participants evaluated them as inadequately describing good design. From this procedure, the design descriptive items courageous, connecting, integrating, precious, stylish were discarded; analogously, the emotion items aroused, amused, confident, desire, energized, anticipating were discarded.

According to Tabachnick and Fidell (2006, 608), PCA aims to 'reduce a large number of observed variables to a smaller number of factors' and 'test a theory about the nature of underlying processes'.

Table 1. Factor structure (PCA).

	Factors			
Items	*Tool*	*Novelty*	*Simplicity*	*Energy*
practical	.41			
functional	.38			
useful	.33			
exciting		.42		
creative		.36		
unique		.36		
simple			.46	
clear			.41	
minimalistic			.38	
powerful				.54
clever				.35
intuitive				.25
Proportion Variance	21%	20%	17%	16%
Cumulative Variance	21%	41%	58%	74%

Note: Factor loadings <.2 are suppressed.

PCA finds the dimensions of greatest variance in a data set and allocates coordinates for each observation along these dimensions. The oblimin rotation method was applied because it accounts for correlations between components, and psychological components tend to be correlated. Each component forms a scale of items which is evaluated in its reliability by the internal consistency metric Cronbach alpha (Cronbach 1951).

Results

The item pool was reduced in subsequent PCA iterations from 32 to 12 design descriptive and from 16 to 5 emotion items, resulting in 17 items (35%) from the original 48.

The final selection of emotions encompassed the items: excited, inspired, pleased, interested, pleasantly surprised.

The design descriptive items showed consistently good psychometric results. With four components extracted, principal component analysis on the final 12 design descriptive items rendered a clear factor structure (see Table 1) – all items show sufficiently high convergent validity (items for one construct load on the same component) and discriminant validity (items for different constructs load on different components and without double loadings). The four components (Table 2) explain the major part of the variance in the survey data (74%). They all show Cronbach alphas near or equal to 0.8, indicating high reliability according to the convention by P. Cohen (1977). The factor correlations (see Table 2) reveal that the Tool factor correlates with all other factors by a medium effect size (.29–.34), whereas the highest correlation is revealed between Novelty and Simplicity ($r = .52$).

Table 2. Reliability and factors correlations.

Factors	Cronbach alpha	*Tool*	*Novelty*	*Simplicity*	*Energy*
Tool	*.80*	1	.34	.29	.34
Novelty	*.80*	.34	1	.52	.13
Simplicity	*.77*	.29	.34	1	.11
Energy	*.76*	.29	.13	.11	1

Beyond the clear factor structure, the most important criterion of exploratory factor analysis is whether a plausible interpretation for all components can be found. The components can be interpreted with sufficiently high face validity and content validity by the following taxonomy:

Tool. The items practical, functional, useful relate to design work perceived as a tool to fulfil a user need, thus denoting solution-oriented aspects.

Novelty. The items exciting, unique, creative relate to design as novelty to the subjective perceiver.

Energy. The items powerful, clever, intuitive tap into a perceived notion of energy, thus representing colloquial terms of intention and intuition.

Simplicity. The items simple, clear, minimalistic denote the aspect of perceived simplicity.

Study 2

Study 2 validated Study 1's measurement instrument for non-expert design perception in a real-world context. Furthermore, its items and dimensions were used in hypothesis testing to predict design preference to provide a partial answer to RQ2 (How well does non-expert design vocabulary predict design preference?) with linear models.

Method

Study 2 evaluated the design preference of design works in three real-world design projects of a brand design consultancy (see Figure 1). These evaluations allowed to analyse how well Study 1's four design dimensions could predict design preference.

Participants

Study 2 acquired a sample of 1955 online survey responses in weekly surveys in three real-world design projects during three months. Students of a graduate and an undergraduate design class at Hongik University, Seoul, South Korea (5 male, 12 female, mean age = 25.5 years, SD = 4.8 years) were assigned to create weekly design deliverables which were shown in an online survey for design evaluations. Raters were the students participating in the design projects as well as three design professors.

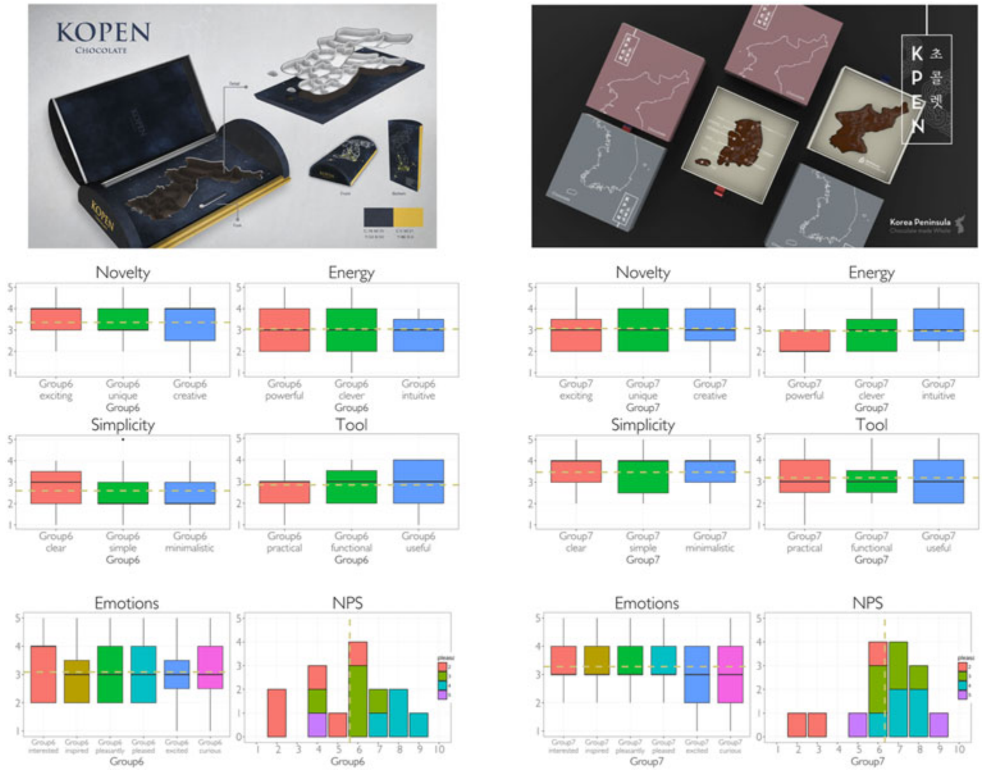

Figure 1.
Feedback analysis in design dimensions.

Measures

Participants evaluated each design work with the design descriptive and emotion items of Study 1 on a 5-point Likert scale (not at all – extremely well).

As dependent variable to measure design preference, the present work applied a derivative of the Net Promoter Score (Reichheld 2003). To yield a metric unbiased from a specific theory, the raw NPS ratings instead of the NPS formula were used. The NPS item was formulated as: ‘How likely would you recommend this design to a friend or colleague as an example of good design?’.

Results

Linear regression per item

All items. When using both the design descriptive and the emotion items simultaneously (see Figure 2), a distinct pattern emerged: All emotion items showed medium-sized regression coefficients (between .27 and .43), larger than the entirety of 12 design descriptive items (between −.06 and .16).

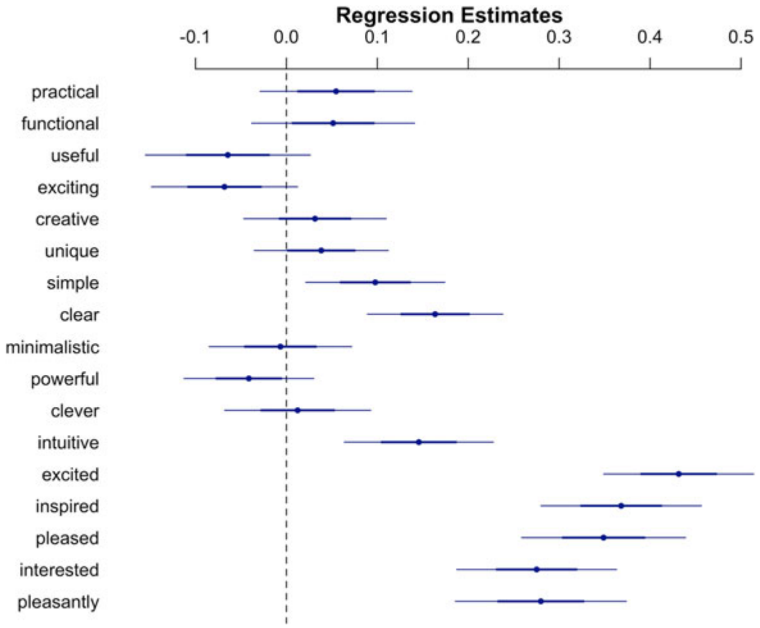

Figure 2.
Linear regression of NPS on all items. pleasantly: pleasantly surprised.

Design descriptive items. A detailed view on the design descriptive items (as only predictors) revealed the regression coefficients to be small (around .2), with the largest values for the items practical (Tool), creative (Novelty), clear (Simplicity), and clever (Energy).

A notable pattern was that high and low regression items were spread evenly across the components. Therefore, it became interesting to investigate the design descriptive impact on design preference on the aggregated component level.

Emotion items. A detailed view of the emotion items as only predictors revealed the highest regression coefficients (of medium to high effect size) for the emotions pleased (.44), excited (.43), and inspired (.40). The emotion inspired denotes the positive influence of creativity on mood. The emotion excited indicates a higher level of arousal. In this vein, the lower regression coefficients of the emotions interested and pleasantly surprised might be explained by their lower arousal level.

Linear regression per dimension

When aggregating the design descriptive items as composite scores on the component level, they showed medium effect sizes for Novelty (.52) and Tool (.48), closely followed by Simplicity (.42) and Energy (.37).

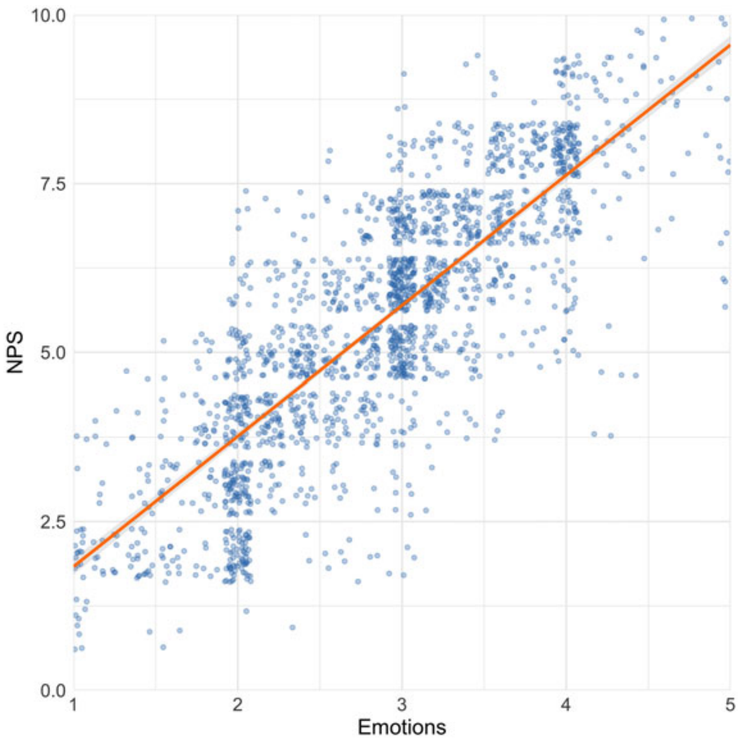

Figure 3.
Linear regression of NPS on Emotions

Emotions, aggregated to a composite score, revealed a beta coefficient on the NPS of .82 on p = .000 (t = 64.53), a large effect size according to the convention of Cohen (1977). The relationship is explained across the whole range of both Emotions and the Net Promoter Score (see Figure 3). It can thus be concluded that emotions serve as a strong and reliable predictor of design preference.

Study 3

The promising results of Study 2 generated the need for validation by other models. Therefore, the goal of Study 3 was to compare the prediction performance of Study 2 with models derived from artificial intelligence.

Method

Study 3 applied machine learning algorithms to predict the Net Promoter Score (NPS) with the design descriptive and emotion items as features.

For supervised learning (Hastie, Tibshirani, and Friedman 2009), neural networks, gradient boosting machines, and random forests are among the best performing models in applied research.

Neural networks. Originally introduced by McCulloch and Pitts (1943), neural networks loosely model the human brain in several layers of nodes that represent a feed-forward mechanism of modeling input data (Rojas 2013). Nodes can have multiple interconnections with other nodes, and each interconnection receives a weight that multiplies the incoming data and sends the sum of weighted inputs to the outgoing connections. The weights are continually adjusted during training until prediction on training data converges.

Gradient boosting machines. Commonly applied to decision trees, gradient boosting machines (GBM) use a specific method – boosting – to improve prediction performance. Boosting assigns higher priority to incorrectly predicted data in the subsequent training interval.

Random forests. The generation of multiple, fully grown decision trees are called random forests. Each tree is grown on a different random draw from the sample (bagging) to maximize the variance between the trees. The prediction estimate is calculated as a majority vote of all grown trees to minimize the bias in the estimation.

Benchmarking machine learning methods

The prediction performance of machine learning models for regression, i.e. prediction of a numeric variable like the NPS, is commonly measured by the root mean squared error (RMSE). The RMSE is defined as the average difference between predicted and observed values. For neural networks, additional parameters (number of layers, number of nodes per layers) were optimized.

The current study performed 10-fold cross-validation in three repetitions for each feature set (design descriptive, emotion, or all items). The training data was centred and scaled within each cross-validation fold. This procedure yielded the RMSE mean and standard deviation as training set performance.

As final validation step, each model's prediction performance was evaluated on the testing set as final benchmarking reference. The testing set had been created by a randomized draw of 25% from the original data.

For comparing predictors on the item level, the present study used 'variable importance', a metric that indicates a variable's relative usefulness in prediction. Variable importance is defined as the mean decrease of accuracy in predictions when this variable is excluded from the model (James et al. 2013, 330).

Results

Study 3's evaluation of the machine learning models allows to benchmark their prediction performances with Study 2's linear regression results.

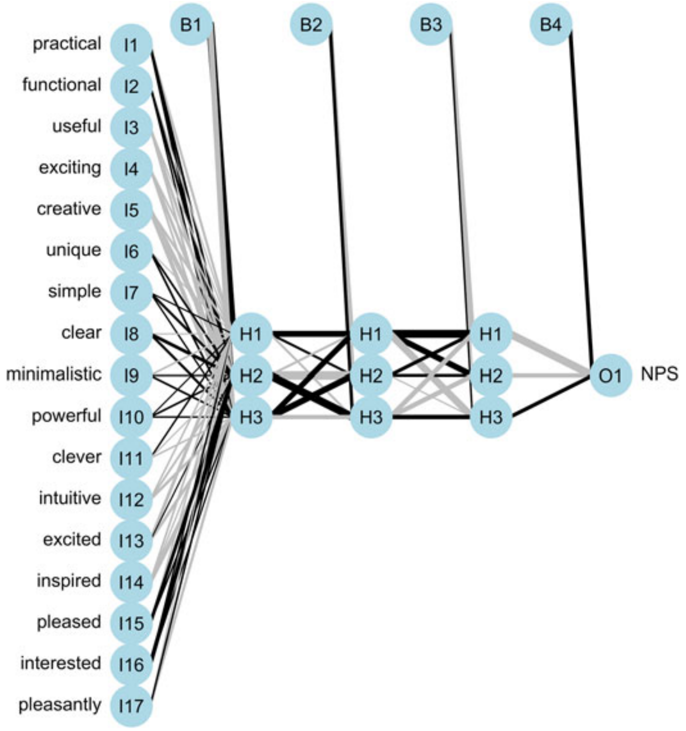

Figure 4.
Neural network 3-layer 3-node architecture.

Table 3. Training set performance benchmarking.

Training Set Performance	*RMSE (10-fold cross-validation, 3 repetitions)*					
Items	*all items*		*design descriptive*		*emotion*	
	mean	*sd*	*mean*	*sd*	*mean*	*sd*
Linear Regression	.557	.026	.765	.059	.559	.041
Neural Networks	.547	.037	.771	.063	.541	.032
Gradient Boosting Machines	.519	.031	.739	.055	.533	.037
Random Forests	.418	.033	.622	.045	.517	.033

Neural networks

With the sample data, a moderately deep network of three layers and eight nodes per layer showed a slightly worse prediction performance as linear regression when using all 17 items. One cause for the bad performance can be 'overfitting', i.e. the model optimizes too well to the training data and consequently underperforms on the validation data. Overfitting is caused by an oversized model capacity which is defined by the number of hidden layers and the number of nodes per layer. Therefore, the present study reduced overfitting by

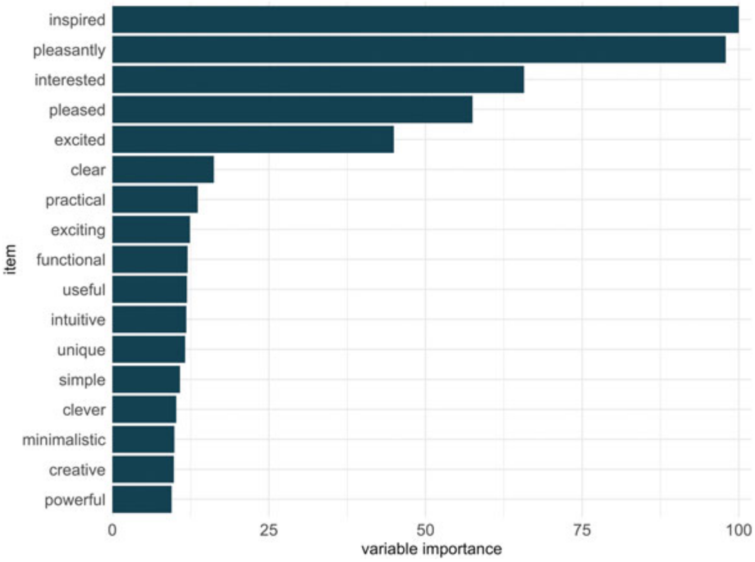

Figure 5.
Variable importance of predictors in random forests. pleasantly = pleasantly surprised

decreasing the model capacity step by step. From all neural networks undergoing this systematic complexity reduction, the best prediction was achieved by a neural network architecture of three layers with 3-3-3 nodes. With this architecture (see Figure 4), the prediction performance of neural networks became nearly identical to linear regression (see Table 3) – showing RMSE differences of less than half a standard deviation with all items ($RMSE_{neural} = .547$ vs. $RMSE_{linear} = .557$), design items ($RMSE_{neural} = .771$ vs. $RMSE_{linear} = .765$) and emotion items ($RMSE_{neural} = .541$ vs. $RMSE_{linear} = .559$).

Gradient boosting machines

Among all items as predictors for gradient boosting machines, all five emotion items showed a variable importance above 50.0% with the highest variable importance for the items inspired (100%) and interested (70.7%), whereas all the design descriptive items remained below 11%. Among only the design descriptive items as predictors, the highest variable importance was shown by the items practical (100%) and creative (65.4%). The gradient boosting machines performed nearly identical to neural networks with design descriptive or emotion items, yet better with all items by nearly one standard deviation ($RMSE_{gbm\text{-}all} = .519$ vs. $RMSE_{neural\text{-}all} = .547$, $sd_{pooled} = .034$).

Random forests

With all items as predictors, random forests showed a similar pattern to gradient boosting machines on variable importance (see Figure 5):

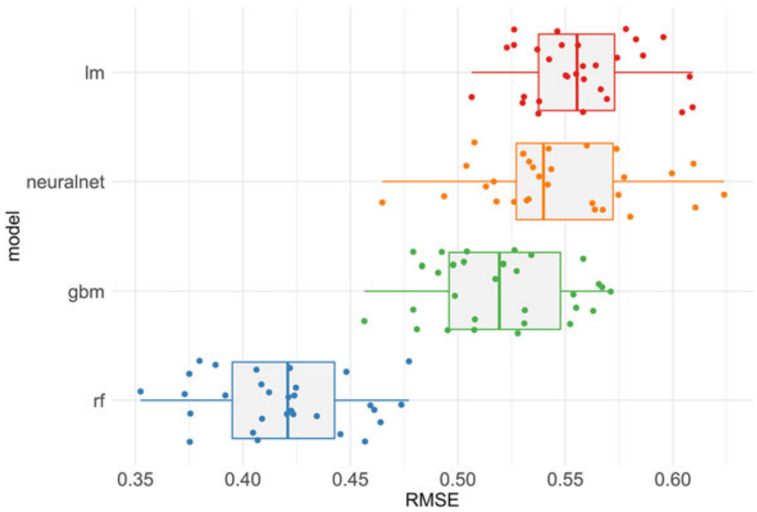

Figure 6.
Sampling distributions of cross-validated training performance.

All emotion items revealed a clearly higher impact on prediction than the design descriptive items; the highest variable importance was generated by the emotion items inspired (100%), pleasantly surprised (97.9%) and interested (65.7%); among the design descriptive items, the highest variable importance was shown again by the items practical (100%) and creative (55.7%).

All random forest models could clearly outperform any other preceding model ($RMSE_{rf\text{-}all}$ = .418, $RMSE_{rf\text{-}design}$ = .622, $RMSE_{rf\text{-}emotion}$ = -.517).

Benchmarking machine learning methods

The visual comparison of the training set performance (Table 3) shows a clear ranking of the models (see Figure 6): on average, all machine learning models performed better than linear regression, as can be seen by the median (middle vertical line) of the depicted boxplots. Gradient boosting machines performed better than neural networks, which in turn performed better than linear regression – however, the mean differences were smaller than one standard deviation and therefore negligible. The clear benchmarking winner is random forests, outperforming gradient boosting machines by more than three pooled standard deviations and linear regression by more than four standard deviations.

The final benchmark was performed on the testing data (see Table 4).

The linear regression's performance was distinctly better when using all items compared to only design descriptive items ($RMSE_{all}$ = .531 vs. $RMSE_{design}$ =.709) but the difference to emotion items was relatively small ($RMSE_{all}$ = .531 vs. $RMSE_{emotion}$ = .586).

Table 4. Testing set performance benchmarking.

Testing Set Performance	*RMSE (75/25 training/testing split)*		
Items	*all items*	*design descriptive*	*emotion*
Linear Regression	.531	.709	.586
Neural Networks	.533	.737	.603
Gradient Boosting Machines	.511	.696	.595
Random Forests	.379	.601	.590

On the testing data, neural networks performed similarly to linear regression or slightly worse. Gradient boosting machines performed slightly better than neural networks for all item sets. Random forests outperformed any other model including gradient boosting machines for all item sets ($RMSE_{rf\text{-}all} = .379$, $RMSE_{rf\text{-}design} = .601$, $RMSE_{rf\text{-}emotion} = .590$).

Taken together, random forests consistently showed the best prediction performance. Nevertheless, when using only emotion items, the linear regression model revealed the same performance as random forests ($RMSE_{lm\text{-}emotion} = .586$ vs. $RMSE_{rf\text{-}emotion} = .590$). This surprising result corroborated Study 2's finding that the relationship between emotions and design preference is strongly linear (β=.82, p = .000).

Discussion

The present work aimed at finding a generalizable model of what defines good design from a non-expert perspective. In this endeavour, Study 1 extracted a minimal design vocabulary of 12 design descriptive and five emotion items from a ranking of 115 items by non-experts. Exploratory factor analysis retrieved a four-factor model of design dimensions. These dimensions could be separated from another factor, Emotions elicited by design.

In Study 2, the four dimensions proved to be good predictors of design preference. With these design dimensions detected and validated, both studies answer RQ1 (How do non-experts perceive design?).

In partial response to RQ2 (How well does non-expert design vocabulary predict design preference?), Study 2 found that the two design dimensions Novelty and Tool are medium-sized predictors of design preference. Moreover, it revealed that Emotions as composite were an almost perfect linear predictor of design preference.

Study 3 corroborated these linear model results with three machine learning methods. These methods again showed that emotions on the item level consistently performed better as predictors than any design descriptive item, answering RQ2.

In summary, the present work produced and validated a generalized model of non-expert design description that effectively measures design perception and predicts design preference.

Theoretical implications

Although preceding research has also produced perception-based measurement instruments, the present work fills two relevant gaps in this literature:

First, the present work contributes a clear insight about users' emotional response compared to other design perceptions: to what degree people like a design product is by far better explained by the emotions elicited than by any other psychological dimension perceived. In other words, there is a distinct and measurable prediction improvement from saying 'this design *is* exciting' to 'this design *makes me feel* excited'.

Secondly, the current work validated this insight on a previously unseen detail level. The additional analyses by machine learning algorithms did not only corroborate the linear model. They also revealed that the emotions inspired, pleasantly surprised and interested predict design preference far better than any other design descriptive item. This detailed item-level analysis has not been shown by other design research efforts to date.

Practical implications

The present work suggests to designers a new way to implement a user-centred design thinking approach. The goal of this approach is to bridge the communication gap between designers and their users because designers usually do not possess knowledge about their users' design vocabulary. With the built measurement instrument, designers can solicit feedback from their users to effectively capture their design perception and emotions elicited by the design work. As the 17 items proved to be good predictors of design preference, designers can receive effective feedback to improve their design. This feedback principle can be implemented by the following design process:

After designers have completed a design prototype, they present it to users by images in a survey. This survey further solicits participants to answer the measurement instrument provided by Study 1. The users' survey responses are analysed and shown back to the designers in a similar fashion as in Figure 1. The designers must then reflect on whether their scores on the emotions elicited are satisfactory or unsatisfactory. The need for improvement is indicated by the lowest-scoring design dimension. This insight defines the design goal of the next prototype development. After completion, a new design iteration starts that repeats the preceding steps.

Going through this design process in several iterations will inevitably make designers understand their users' perspective better. Moreover, their design efforts become focused on improving their users' design preferences rather than their own. This shift of focus may resolve occasional design fixations when designers become too focused on their own ideas.

Relevance

Instead of online systems that represent a product, the present work focuses on products that are displayed in online systems. This different focus yields more relevance to the findings for designers because they are increasingly working and displaying their design work online. As a result, people are more likely to see design work in social media, so their first interaction with the design work is with its online display. In this scenario, they are free from any preconception originating from personal attitudes to products or brands. Furthermore, consumers increasingly search for product recommendations online and thus obtain the first impression of a product design in an online system.

Conclusions

In conclusion, this study contributes a new participatory design process to the design thinking practice by reducing the semantic barrier between non-experts and designers. The presented measurement instrument can effectively improve a feedback process from users to designers as it provides a design vocabulary that non-experts can understand and that predicts their design preferences.

Acknowledgments

The author would like to thank Alex Hayes (University of Wisconsin-Madison) for his helpful comments on evaluation of machine learning models.

Funding

This work was supported by the 2016 Hongik University New Faculty Research Fund.

References

Abras, C., D. Maloney-Krichmar, and J. Preece. 2004. "User-Centered Design." In *Encyclopedia of Human-Computer Interaction*., edited by W. Brainbridge, 37, 445–456. Thousand Oaks, CA: Sage Publications.

Ann, M., and M. A. N. N. Stankiewicz. 1987. "Beauty in Design and Pictures: Idealism and Aesthetic Education." *Journal of Aesthetic Education* 21 (4): 63–76. doi:10.2307/3332831.

Bangor, A., P. Kortum, and J. Miller. 2009. “Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale.” *Journal of Usability Studies* 4 (3): 114–123. https://doi.org/66.39.39.113.

Bangor, A., P. T. Kortum, and J. T. Miller. 2008. “An Empirical Evaluation of the System Usability Scale.” *International Journal of Human-Computer Interaction* 24 (6): 574–594. doi:10.1080/10447310802205776.

Biskjaer, M. M., and K. Halskov. 2014. “Decisive Constraints as a Creative Resource in Interaction Design.” *Digital Creativity* 25 (1): 27–61. doi:10.1080/14626268.2013.855239.

Bostrom, A. 1997. “Risk Perceptions: ‘Experts’ vs. ‘Lay People.” *Duke Environmental Law & Policy Forum* 8 (1): 101–113. https://doi.org/10.3366/ajicl.2011.0005.

Büttner, S., and C. Röcker. 2016. “Applying Human - Centered Design Methods in Industry – a Field Report.”

Card, S. K. 2017. *The Psychology of Human-Computer Interaction*. Boca Raton, FL: CRC Press.

Chamberlain, A., T. Rodden, M. Jones, S. Park, and Y. Rogers. 2012. “Research in the Wild : Understanding ‘In the Wild ’ Approaches to Design and Development.” *Proceedings of the Designing Interactive Systems Conference*, 795–796. https://doi.org/10.1145/2317956.2318078.

Chi, M. 2006. “Methods to Assess the Representations of Experts’ and Novices’ Knowledge.” In *Cambridge Handbook of Expertise and Expert Performance*., edited by K. A. Ericson, N. Charson, and P. Feltovich, 167–184. Cambridge: Cambridge University Press.

Cohen, P. 1977. *Statistical Power Analysis for the Behavioral Sciences*. 1st ed. New York: NY: Academic Press.

Cohen, R. J., M. E. Swerdlik, and S. M. Phillips. 1996. *Psychological Testing and Assessment: An Introduction to Tests and Measurement*. Mountain View, CA: Mayfield Publishing.

Cook, G., E. Pieri, and P. T. Robbins. 2004. “The Scientists Think and the Public Feels’: Expert Perceptions of the Discourse of GM Food.” *Discourse & Society* 15 (4): 433–449. doi:10.1177/0957926504043708.

Cronbach, L. J. 1951. “Coefficient Alpha and the Internal Structure of Tests.” *Psychometrika* 16 (3):297–334. doi:10.1007/BF02310555.

Desmet, P. M. A. 2012. “Faces of Product Pleasure: 25 Positive Emotions in Human-Product Interactions.” *International Journal of Design* 6 (2): 1–29.

Desmet, P., K. Overbeeke, and S. Tax. 2001. “Designing Products with Added Emotional Value: Development and Application of an Approach for Research through Design.” *The Design Journal* 4 (1): 32–47. (April): doi:10.2752/146069201789378496.

Fisher, C. D. 2003. “Why Do Lay People Believe That Satisfaction and Performance Are Correlated? Possible Sources of a

Commonsense Theory." *Journal of Organizational Behavior* 24 (6): 753–777. doi:10.1002/job.219.

Flores-Mir, C., E. Silva, M. I. Barriga, M. O. Lagravère, and P. W. Major. 2004. "Lay Person's Perception of Smile Aesthetics in Dental and Facial Views." *Journal of Orthodontics* 31 (3): 204–209. doi:10.1179/146531204225022416.

Fredrickson, B. L. 2001. "The Role of Positive Emotions in Positive Psychology. The Broaden-and-Build Theory of Positive Emotions." *American Psychologist* 56 (3): 218–226. doi:10.1037/0003-066X.56.3.218.

Friestad, M., and P. Wright. 1995. "Persuasion Knowledge: Lay People's and Researchers' Beliefs about the Psychology of Advertising." *Journal of Consumer Research* 22 (1): 62–74. doi: 10.1086/209435.

Galiana, M., C. Llinares, and Á. Page. 2012. "Subjective Evaluation of Music Hall Acoustics: Response of Expert and Non-Expert Users." *Building and Environment* 58: 1–13. (December): doi: 10.1016/j.buildenv.2012.06.008.

Ghatak, A. 2017. *Machine Learning with R*. *Machine Learning with R.* https://doi.org/10.1007/978-981-10-6808-9.

Giles, D. C. 2002. *Advanced Research Methods in Psychology*. East Sussex, UK and New York: Routledge.

Gredler, M. E. 1997. *Learning and Instruction: Theory into Practice*. 3rd ed. Upper Saddle River, NJ: Prentice Hall.

Hanington, B. 2003. "Methods in the Making: A Perspective on the State of Human Research in Design." *Design Issues* 19 (4): 9–18. doi:10.1162/074793603322545019.

Hassenzahl, M. 2001. "The Effect of Perceived Hedonic Quality on Product Appealingness." *International Journal of Human-Computer Interaction* 13 (4): 481–499. doi:10.1207/S15327590IJHC1304_07.

Hassenzahl, M. 2004. "The Interplay of Beauty, Goodness, and Usability in Interactive Products." *Human Computer Interaction* 19 (4): 319–349. https://doi.org/10.1207/s15327051hci1904.

Hassenzahl, M., M. Burmester, and F. Koller. 2003. "AttrakDiff : Ein Fragebogen Zur Messung Wahrgenommener Hedonischer Und Pragmatischer Qualität." *Mensch & Computer 2003. Interaktion in Bewegung*: 187–196. https://doi.org/10.1007/978-3-322-80058-9.

Hassenzahl, M., and N. Tractinsky. 2006. "User Experience - A Research Agenda." *Behaviour & Information Technology* 25 (2): 91–97. doi:10.1080/01449290500330331.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. "Overview of Supervised Learning." In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*., edited by Trevor Hastie, Robert Tibshirani, and Jerome Friedman, 9–41. New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-84858-7_2.

Hekkert, P., and P. C. W. Van Wieringen. 1996. “Beauty in the Eye of Expert and Nonexpert Beholders: A Study in the Appraisal of Art.” *American Journal of Psychology* 109 (3): 389–407. doi: 10.2307/1423013.

Hodson, G., and V. M. Esses. 2005. “Lay Perceptions of Ethnic Prejudice: Causes, Solutions, and Individual Differences.” *European Journal of Social Psychology* 35 (3): 329–344. doi: 10.1002/ejsp.251.

Hollan, J. D. J., E. Hutchins, and D. Kirsh. 2000. “Distributed Cognition: Toward a New Foundation for Human-Computer Interaction Research.” *ACM Transactions on Computer-Human Interaction* 7 (2): 174–196. doi:10.1145/353485.353487.

Huang, G. B., M. Mattar, T. Berg, and E. Learned-Miller. 2007. “Labeled Faces in the Wild: A Database ForStudying Face Recognition in Unconstrained Environments.”

Hutchins, E. 1995. *Cognition in the Wild*. Cambridge, MA: MIT Press.

Jacko, J. A. 2012. *Human Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications*. Edited by Julie A. Jacko. Boca Raton, FL: CRC Press.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An Introduction to Statistical Learning*. New York: Springer.

Ju, W., and L. Leifer. 2008. “The Design of Implicit Interactions: Making Interactive Systems Less Obnoxious.” *Design Issues* 24 (3): 72–84. doi:10.1162/desi.2008.24.3.72.

Karvonen, K. 2000. “The Beauty of Simplicity.” *Proceedings on the 2000 Conference on Universal Usability - CUU ’00*, 85–90. https://doi.org/10.1145/355460.355478.

Keiningham, T. L., L. Aksoy, and B. Cooil. 2008. “Linking Customer Loyalty to Growth.” *MIT Sloan Management Review* (49414): 51–57. https://doi.org/10.1016/j.ymssp.2004.04.003.

Keiningham, T. L., L. Aksoy, B. Cooil, T. W. Andreassen, and L. Williams. 2008. “A Holistic Examination of Net Promoter.” *Journal of Database Marketing & Customer Strategy Management* 15 (2): 79–90. doi:10.1057/dbm.2008.4.

Kline, P. 1998. *The New Psychometrics: Science, Psychology, and Measurement*. London: Routledge.

Kuhn, M., and K. Johnson. 2013. *Applied Predictive Modeling*. Vol. 26. New York: Springer.

Li, W., F. Abtahi, C. Tsangouri, and Z. Zhu. 2016. “Towards an ‘In-the-Wild’ Emotion Dataset Using a Game-Based Framework.” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 1526–1534. https://doi.org/10.1109/CVPRW.2016.190.

Mano, H., and R. Oliver. 1993. “Assessing the Dimensionality and Structure of the Consumption Experience: Evaluation, Feeling, and Satisfaction.” *Journal of Consumer Research* 20 (3): 451. doi: 10.1086/209361.

Marsh, T., W. L. Wong, and E. Carriazo. 2005. “User Experiences and Lessons Learned from Developing and Implementing an Immersive Game for the Science Classroom.” In *Proceedings of HCI International 2005*.

McCulloch, W. S., and W. Pitts. 1943. “A Logical Calculus of the Idea Immanent in Nervous Activity.” The Bulletin of Mathematical Biophysics 5 (4): 115–133. doi:10.1007/BF02478259.

Minh, T., T. Do, J. Blom, and D. Gatica-Perez. 2011. “Smartphone Usage in the Wild: A Large-Scale Analysis of Applications and Context.” *ICMI ’11 Proceedings of the 13th International Conference on Multimodal Interfaces*, 353–360. https://doi.org/10.1145/2070481.2070550.

Norman, D. A., and S. W. Draper. 1986. *User-Centered Systems Design: New Perspectives on Human-Computer Interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Reichheld, F. F. 2003. “The One Number You Need to Grow.” *Harvard Business Review* 81 (12):46–54.

Rojas, R. 2013. *Neural Networks: A Systematic Introduction*. Chicago, IL: Springer Science & Business Media.

Rosson, M. B., and J. M. Carroll. 2002. *Usability Engineering: Scenario-Based Development of Human-Computer Interaction*, edited by Diane D. Cerra. San Francisco, CA: Morgan Kaufmann.

Sakalaki, M., and S. Kazi. 2007. “How Much Is Information Worth? Willingness to Pay for Expert and Non-Expert Informational Goods Compared to Material Goods in Lay Economic Thinking.” *Journal of Information Science* 33 (3): 315–325. doi:10.1177/0165551506070709.

Sharit, J. 2006. *Handbook of Human Factors and Ergonomics*. 3rd ed. Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/0470048204.ch27.

Snow, R., B. O’Connor, D. Jurafsky, and A. Y. Ng. Dolores Labs, and Capp St. 2008. “Cheap and Fast - but Is It Good? Evaluation Non-Expert Annotations for Natural Language Tasks.” Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), no. October: 254–263. https://doi.org/10.1.1.142.8286.

Sonderegger, A., and J. Sauer. 2010. “The Influence of Design Aesthetics in Usability Testing: Effects on User Performance and Perceived Usability.” *Applied Ergonomics* 41 (3): 403–410. doi:10.1016/j.apergo.2009.09.002.

Tabachnick, B. G., and L. S. Fidell. 2006. *Using Multivariate Statistics*. 5th ed. Boston, MA: Pearson Education.

Tuch, A. N., S. P. Roth, K. Hornbaek, K. Opwis, and J. A. Bargas-Avila. 2012. “Is Beautiful Really Usable? toward Understanding the Relation between Usability, Aesthetics, and Affect in HCI.” *Computers in Human Behavior* 28 (5): 1596–1607. doi:10.1016/j.chb.2012.03.024.

Turner, S. 2001. “What Is the Problem with Experts?” *Social Studies of Science* 31 (1): 123–149. doi:10.1177/030631201031001007.

Westbrook, R. A., and R. Oliver. 1991. “The Dimensionality of Consumption Emotion Patterns and Consumer Satisfaction.” *Journal of Consumer Research* 18 (1): 84–91. (October): doi: 10.1086/209243.

Wickens, C. D., S. E. Gordon, and Y. Liu. 2004. *An Introduction to Human Factors Engineering*. London: Pearson.

Yilmaz, S., and C. M. Seifert. 2011. “Creativity through Design Heuristics: A Case Study of Expert Product Design.” *Design Studies* 32 (4): 384–415. doi:10.1016/j.destud.2011.01.003.

Yilmaz, S., and C. Seifert. 2009. “Cognitive Heuristics Employed by Designers.” *Design Science*: 4243–4246.

Yoon, J. K., P. M. A. Desmet, and A. van der Helm. 2012. “Design for Interest: Exploratory Study on a Distinct Positive Emotion in Human-Product Interaction.” *International Journal of Design* 6 (2): 67–80.

Zhu, X., and, D. Ramanan. 2012. “Face Detection, Pose Estimation, and Landmark Estimation in the Wild.” *Cvpr*: 2879–2886. https://doi.org/10.1109/CVPR.2012.6248014.

Zimbardo, P. G., R. L. Johnson, and V. McCann. 2012. *Psychology Core Concepts. Always Learning*. London: Pearson. https://doi.org/10.1016/S1053-8119(05)70016-1.

Biography

Chaehan So is a Professor of Design Psychology at the International School of Design for Advanced Studies (IDAS), Hongik University, in Seoul. His research interests are in design thinking, experience design, and applied artificial intelligence. Dr. So is merging these methods into design psychology – new and psychologically-grounded ways of creating user experience. He holds a Ph.D. in Psychology (Humboldt University of Berlin, Germany), an M.S. in business (Ecole Supérieure de Commerce de Paris, France) and an M.S. in electrical engineering (Technical University Berlin, Germany).

ORCID

Chaehan So http://orcid.org/0000-0002-0546-2947

Address for Correspondence

Chaehan So, Department of Smart Design Engineering & Department of Artificial Intelligence/Big Data, International Design School for Advanced Studies, Hongik University, 57 Daehak-ro, Jongno-gu, 03082 Seoul, South Korea. Tel: +82 2 3668 3825, Email: cso@hongik.ac.kr