Establishing meaningful cut points for online user ratings

Gerrit Hirschfeld[1] and Meinald T. Thielsch[2]

[1] German Paediatric Pain Centre, Children's Hospital Datteln, Dr.-Friedrich-Steiner Str. 5, 45711 Datteln, Germany
[2]Department of Psychology, University of Münster, Fliednerstr. 21, 48149 Münster, Germany

Address for correspondence:
Gerrit Hirschfeld; German Paediatric Pain Centre, Children's and Adolescents' Hospital, Datteln, Dr.-Friedrich-Steiner Str. 5, 45711 Datteln, Germany; Tel.+49- 2363-975-183; Fax.+49- 2363-975-181; e-mail: g.hirschfeld@deutsches-kinderschmerzzentrum.de

Subjective perceptions of websites can be reliably measured with questionnaires. But it is unclear how such scores should be interpreted in practice, e.g. is an aesthetics score of 4 points on a 7-point-scale satisfactory? The current paper introduces a ROC-based methodology to establish meaningful cut points for the VisAWI (Visual Aesthetics of Websites Inventory) and it's short form the VisAWI-S. In two studies we use users' global ratings (UGRs) and website rankings as anchors. A total of 972 participants took part in the studies and yielded similar results. First, one-item UGRs correlate highly with the VisAWI. Second, cut points on the VisAWI reliably differentiate between sites that are perceived as attractive vs. unattractive. Third, these cut points are variable, but only within a certain range. Together the research presented here establishes a score of 4.5 on the VisAWI is a reasonable goal for website designers and highlight the utility of the ROC methodology to derive relevant scores for rating scales.

Keywords; aesthetics; website evaluation: first impression; global ratings; VisAWI

Practitioner summary

We demonstrate the benefit of ROC-based methods in finding relevant cut points for online user ratings. Specifically, we establish that a score of "4.5" as a meaningful cut point for the VisAWI, a scale for measuring aesthetic appeal of websites, and it's short form, the VisAWI-S.

1. Introduction

The major role the World Wide Web plays in the everyday life of many people leads to questions of how people perceive and evaluate websites. Thus, many studies try to understand the key components of users website perception and aim to derive methods to quantify such online user ratings (e.g. Flavián, Guinalíu, and Gurrea 2006; Hornbæk 2006; Lavie and Tractinsky 2004; Lee and Koubek 2012; Thielsch, Blotenberg, and Jaron 2014). A core aspect in the evaluation of websites is the assessment whether a website is aesthetically pleasing or not. Website aesthetics can be defined as an immediate pleasurable and subjective experience that is directed toward a website and not mediated by intervening reasoning (Moshagen and Thielsch 2010). Aesthetics has been shown to influence amongst others; a) the perception of usability (for an overview see: Lee and Koubek 2012) b) actual performance (Moshagen, Musch, and Göritz 2009; Sonderegger and Sauer 2010) or c) even perceptions of website content (De Angeli, Sutcliffe, and Hartmann 2006; Thielsch, Blotenberg, and Jaron 2014). Aesthetic design seems to shape especially first impressions of a website (Thielsch, Blotenberg, and Jaron 2014; Thielsch and Hirschfeld 2012; Tuch et al. 2012). Thus, the evaluation of aesthetics is an important part of website tests and target for the improvement of existing websites. Such an assessment requires reliable measures for aesthetics, and meaningful standards for their interpretation, i.e. cut points that define relevant levels of aesthetics by specifying which scores from a continuous scale are relevant.

Since aesthetics plays an important role in website appraisal, much effort has been put into developing measures of website aesthetics. The reliability and validity of some measures, e.g the instrument by Lavie and Tractinsky (2004) or the VisAWI (Visual Aesthetics of Websites Inventory, Moshagen and Thielsch 2010), has been established, but there are no agreed-upon cut points that could be used to guide practical decisions. In practice, ratings of specific websites are most often interpreted by relying on direct comparisons between alternatives. Rarely benchmarks are used to interpret the ratings to specific websites. Inspired by methods in medicine (Copay et al. 2007) we suggest to use users' global ratings (UGRs) of first and overall impression or their ranking of websites as anchors against which we evaluate possible cut points for aesthetics by means of receiver-operating characteristic (ROC) methods. This allows defining optimal cut points as those differentiate best between websites that are perceived as generally positive vs. those that are perceived as negative.

Aesthetics lends itself easily to the use of ROC-methods. Besides the named importance of aesthetics, the perception processes behind the construct itself fits very well to the idea of a cut point: The principle of an aesthetic threshold is known in empirical aesthetics almost since the beginning of research in this area. One of the main founders of experimental aesthetics, Gustav Theodor Fechner, discussed this threshold as the first principle of aesthetics (Fechner 1876). Based on his research in psychophysics (Fechner 1860), he differentiated between intensity of a stimulus and excitability of a perceiver. These ideas were picked up later by psychobiological theories of aesthetics (Berlyne 1971; Berlyne 1974) and are now well used in neuroaesthetics (Jacobsen 2006). Our idea is to use this mechanism of an aesthetic stimulus exceeding a certain threshold to evoke aesthetic appreciation of a perceiver in using data from website tests. In doing so, we can calculate which cut point in the user ratings is the threshold to evaluate a website as aesthetic. Before describing the aims of the present study in more detail, we briefly describe measures for website aesthetics and the method to define cut points.

1.1 Measures for website aesthetics

In theory there are several ways to measure visual aesthetics, not only questionnaires but also paired-comparisons tasks, formalized mathematical assessments, or physiological measures (for an overview see Moshagen and Thielsch 2010, p. 691). In practice most previous studies used a questionnaire approach including single-item aesthetics measures, ad-hoc developed scales, single scales only partly measuring aesthetics taken from more general website evaluation instruments, or instruments specifically created to measure website aesthetics. At the moment, there are two very well designed and validated instruments for measuring the visual aesthetics of websites available: the instrument of Lavie and Tractinsky (2004) and the VisAWI (Moshagen and Thielsch 2010).

Lavie and Tractinsky (2004) identified two dimensions of visual website aesthetics; classic aesthetics (containing aspects like symmetry or clearness) and expressive aesthetics (containing aspects like creativity and originality). This instrument was validated using exploratory and confirmatory factor analyses. Convergent and divergent validity was established by demonstrating high correlations to a measure of pleasure and moderate correlations to measures of usability and service quality.
The VisAWI (Moshagen and Thielsch 2010) was created in a series of four studies in which four sub-facets were identified that all load on a general factor: simplicity (partly corresponding to the classical aesthetics of Lavie and Tractinsky 2004), diversity, color, and craftsmanship. In a series of three studies Moshagen and Thielsch (2010) provided evidence for the reliability of the VisAWI as well as for convergent, divergent, discriminative, concurrent and experimental validity. In three additional studies they created a short version of the measure called VisAWI-S and demonstrated its reliability, convergent, divergent, and concurrent validity as well as the strong relation of the measured general aesthetics score to the full VisAWI (Moshagen and Thielsch 2013).

Thus, both instruments allow for a reliable and valid evaluation of users' perceptions of website aesthetics. Additionally, the VisAWI is able to measure a general factor of aesthetic website perception and a four item short version is provided. But in practice, the given ratings on the VisAWI can only be interpreted in an useful manner when different versions of a website are compared. Thielsch and Moshagen (2011) provided benchmark data of 102 tested websites combined in ten categories, but further practical advice or extended benchmarks are still needed. At this point, we would like to introduce several methods other than benchmarks to interpret continuous scales.

1.2 Developing optimal cut points for continuous scales

Many scientific domains have to cope with the fact that tests yield continuous outcomes while practical decisions take a "yes or no" form. This is especially relevant to medicine were treatment decisions often have to be made based on continuous test results. Because of the high relevance of this issue procedures have been developed to determine cut points that are relevant to patients and can be agreed upon within the community. These methods are based on standard ROC methodology (Copay et al. 2007). Since receiver-operating characteristic (ROC)-based methods have to our knowledge not been applied in the domain of website evaluations, we want to introduce the basic terminology, and describe important extension of these methods. ROC-based methods aim to characterize and optimize the performance of diagnostic systems (Swets 1988). If a diagnostic system is used to distinguish between two classes of events - most generally "signal" and "noise" - it's performance can be described by a two by two table. Diagnostic performance is quantified by two indices; sensitivity (ability to correctly identify presence of a signal) and specificity (ability to correctly identify absence of a signal). Critically these methods can be used to rationally decide on the cut point for a continuous scale that best discriminates a binary anchor. This is done by plotting the sensititivity and specificity of all possible cut points against each other. An optimal cut point has at the same time a high level of sensitivity and high level of specificity. The Youden index (sum of sensitivity and specificity minus one) is a common one-number index on which such decisions are based even though resulting cut points are subject to chance variation (Fluss, Faraggi, and Reiser 2005; Schisterman and Perkins 2007). In ergonomics these methods have for example been utilized to fine-tune automatic alarm systems (Bustamante, Bliss, and Anderson 2007).

1.3 Aim of the present research

The aim of the present research was to define relevant cut points for assessing aesthetics with the VisAWI using ROC-based methods. Towards this goal we conducted two studies. The first uses UGRs of first and overall impression as an anchor for optimal cut points on the continuous VisAWI scale. The second uses participants ranking as an anchor for optimal cut points on the VisAWI-S scale. In both studies we use the Youden index as criterion to define optimal cut points. In order to circumvent some problems with optimal cut points, most notably their chance variation, we use bootstrapping to estimate their variability.

2. Study 1

2.1. Method

2.1.1. Participants

A total of 618 participants took part in this study; 321 were female (51.9 %). Ages ranged from 15 to 82 years (*M* = 34.94, *SD* = 13.65). The education level of 78.7 % of the participants was high school diploma or higher. On average, the participants had been using the Internet for 11.66 years (Min = 2, Max = 30, SD = 5.12) and stated an active use of on average 2.52 hours a day (Min = 0.2, Max = 12, SD = 1.92). Participants took part voluntarily and on an anonymous basis.

2.1.2. Stimulus material

A set of 30 websites from ten different content domains (readers are referred to Thielsch and Hirschfeld 2010, for a description of this categorization scheme) was used. These websites were selected to represent a broad range of corporate and institutional websites in Germany, including amongst others corporate websites, e-commerce, e-recruiting, entertainment, and information sites. Each website category was represented by two to five websites.

2.1.3. Measures

For the evaluation of perceived visual aesthetics we used the VisAWI (Moshagen and Thielsch 2010). This questionnaire consists out of 18 items measuring a general aesthetic factor consisting of four facets (“simplicity”, “diversity”, “color”, and “craftsmanship”). Moshagen and Thielsch (2013) created as well a short form, consisting only of four items. Participants were asked to indicate their level of agreement to each VisAWI-item on seven-point Likert scale ranging from 1 (“strongly disagree”) to 7 (“strongly agree”). The authors report a Cronbach’s α of .81 (VisAWI-S) and .94 (full VisAWI), and provided evidence for convergent, divergent, discriminative, concurrent and experimental validity of the VisAWI.
Additionally, participants were asked to rate their first impression (“My first impression: I would mark the website with…”), as well as their overall impression of the website (“Altogether: I would mark the website with…”) on a six-point scale ranging from “insufficient” to “very good” that is used to grade students performance in German the education system from primary to tertiary education settings and as such is well-known by the participants.

2.1.4. Procedure

Participants were recruited via the panel of the German online platform PsyWeb (https://psyweb.uni-muenster.de/). Participation in the panel is completely voluntarily and members agree that there may be invited to studies; they can unsubscribe and delete their personal data at any time. Participants received an e-mail inviting them to participate in a study about the evaluation websites. The e-mail contained a link to the questionnaire via which the data was collected. After being asked for some demographic information (age, gender, education level, Internet experience), participants were randomly assigned to one website from the stimulus set. The given website was presented within a split screen, the questions regarding the website were presented in the smaller upper panel. At the beginning, participants were asked to rate their first impression of the website. Next, they were instructed to explore and to open some subpages of the given website. Then they answered the 18 VisAWI and VisAWI-S items (and two other measures not pertinent to this study). The scales used in the middle part of the study were given in random order, and all items within the questionnaires were also randomized. Afterwards, the overall impression was rated on the same scale used at the beginning. At the end of the study, participants were thanked. They were given the opportunity to exclude their data from the subsequent analysis if they whished and to comment on the study. On average, participants took 10 to 12 minutes to complete the study.

2.1.5. Data analysis

Data were analyzed in three steps. First, we used scatter plots to describe the relation between the VisAWI scores (both the full VisAWI and the VisAWI-S) and users’ global ratings (UGRs). Second, we use a median-split that collapsed the two highest UGR categories vs. the five lowest to yield a dichotomous rating (“good” vs. “bad” websites) against which different cut points for the continuous VisAWI were evaluated. Specifically, we constructed a ROC-curve that displays the sensitivity and specificity for all possible cut points and calculated the Area Under Curve (AUC) value as an index of the overall diagnostic utility independent of a specific cut point. We defined the cut point as optimal that maximized the Youden index (sum of sensitivity and specificity minus one). Third, we use bootstrapping to replicate the analysis in pseudo samples with similar characteristics. Specifically we drew (with replacement) 1,000 pseudo samples from the original population with the same size as the original population and recorded the optimal cut points that result in each. Data analysis was performed in R, the code to reproduce the analysis will be shared upon request.

2.2. Results

2.2.1. Correlations between UGRs and VisAWI ratings

Overall we found a large and highly significant correlation between the VisAWI and the first impression UGRs (r = .69; 95% CI = .65 to .73; p < .001) and the overall impression UGRs (r = .78; 95% CI = .74 to .81; p < .001). The same was true for the VisAWI-S that also correlated highly with the first impression UGRs (r = .75; 95% CI = .71 to .78; p < .001) and the overall impression UGRs (r = .67; 95% CI = .62 to .71; p < .001). As can be seen in figure 1 the means in the different UGR-categories can be fitted very well to a linear regression line.

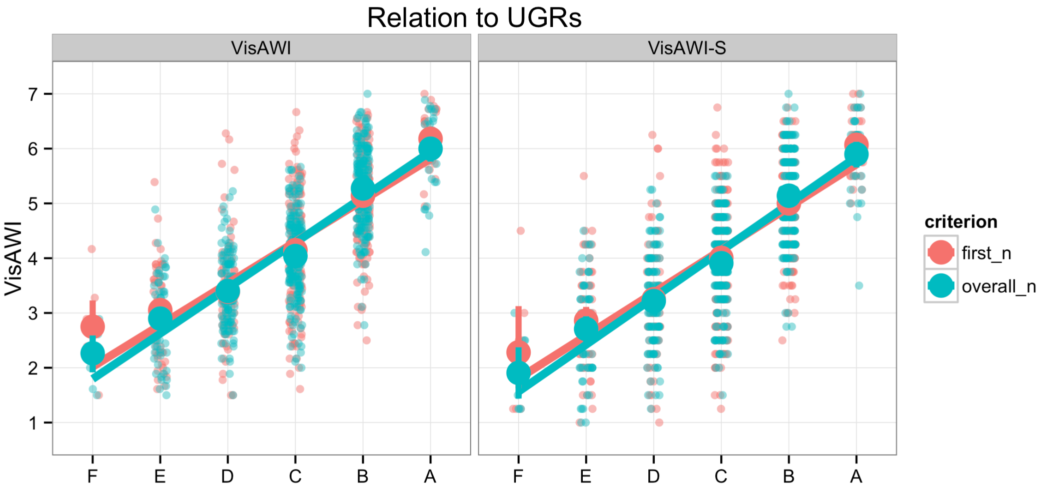

Figure 1. Relation between participants' global impression and VisAWI (Moshagen and Thielsch 2010) scores.
Note: Error bars represent 95% CI for the mean. Straight lines represent best linear fit.

2.2.2. Optimal cut points for the whole sample

Inspection of the ROC-curve showed that the VisAWI scores could be used to distinguish between overall "good" vs. "bad" sites (fig. 2). For the VisAWI the corresponding AUC (Area under curve)-values of .87 (95% CI .84 to .90) for the first impression UGRs and .93 (95% CI .92 to .95) for the overall impression UGRs can be considered "rather high accuracy" according to established standards (Swets 1988). Applying the criteria for optimal cut points as described above yielded a cut point of "4.39" for both the first impression UGRs (sensitivity = .87; specificity = .74) and overall impression UGRs (sensitivity = .92; specificity = .8) criteria.

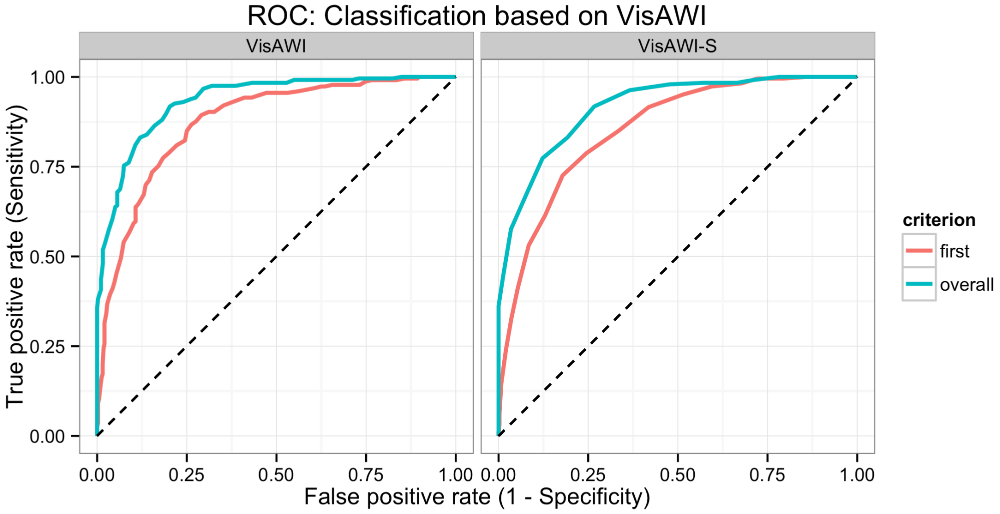

Figure 2. ROC-curve for the VisAWI (Moshagen and Thielsch 2010) against the dichotomous good vs. unattractive rating.
Note: Broken line represent chance classification

Table 1. Performance of various alternative cut points in the full sample

	First impression		Overall impression	
cut point	sensitivity	specificity	sensitivity	specificity
7	0.01	1	0.01	1
6.75	0.03	1	0.03	1
6.5	0.06	1	0.06	1
6.25	0.15	0.99	0.15	1
6	0.24	0.98	0.26	1
5.75	0.33	0.96	0.36	1
5.5	0.41	0.95	0.45	0.99
5.25	0.53	0.92	0.58	0.97
5	0.62	0.87	0.67	0.93
4.75**	0.73	0.82	0.77	0.88
4.5*	0.79	0.76	0.83	0.81
4.25	0.85	0.67	0.92	0.73
4	0.92	0.58	0.96	0.63
3.75	0.95	0.48	0.98	0.52
3.5	0.97	0.41	0.98	0.43
3.25	0.98	0.32	0.98	0.34
3	1	0.28	0.99	0.29
2.75	1	0.21	1	0.22
2.5	1	0.14	1	0.15
2.25	1	0.09	1	0.1
2	1	0.05	1	0.05
1.75	1	0.04	1	0.04
1.5	1	0.02	1	0.02
1.25	1	0.01	1	0.01
1	1	0	1	0

Note. * and ** = Optimal cut point according to first impression* and global impression**.

For the VisAWI-S the AUC-values were slightly lower with .85 (95% CI .82 to .88) and .92 (95% CI .90 to .94). For the VisAWI-S two different cut points emerged for the two UGRs. Specifically, the first impression UGRs as criterion suggested "4.5" (sensitivity = .79; specificity = .76) as cut point while the global impression UGRs as criterion suggested "4.75" as cut point. However, as can be seen in table 1 several alternative cut points yielded similarly high sensitivities and specificities, indicating, that while different cut points emerged, a common alternative may be viable for the different questionnaire versions.

2.2.3. Variability of optimal cut points

Our bootstrapping analysis showed that the different cut points showed some variability (fig. 3). Importantly for both VisAWI versions and both criteria, a cut point of "4.5" was within the 95% CI. We suggest that this should be used as cut point, i.e. VisAWI scores for a specific website lower than "4.5" should be considered bad and values of "4.5" and higher as good.

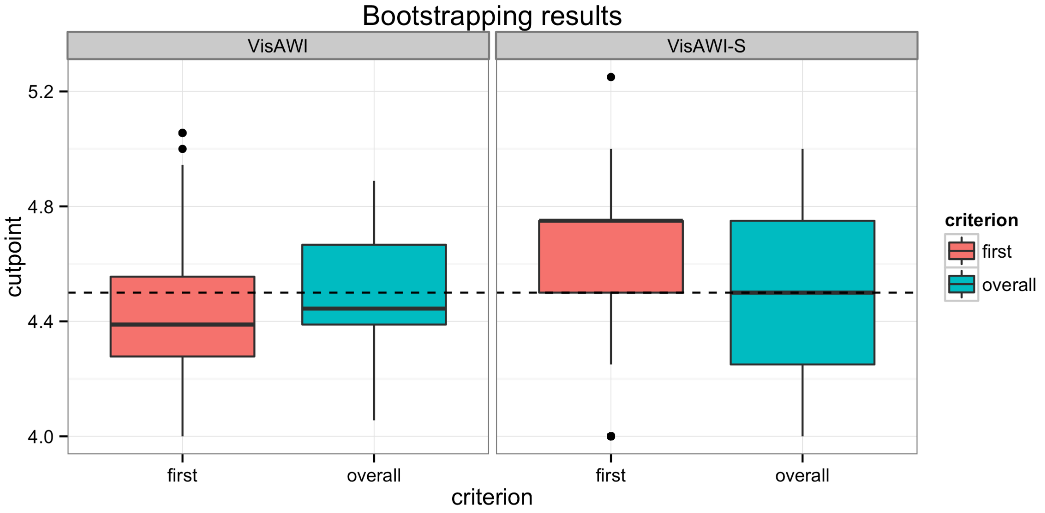

Figure 3. Boxplots for the optimal cut points selected in the bootstrapping-samples. Note: Broken line represent 5% criterion.

2.3. Discussion

The aim of the first study was to define meaningful scores on the VisAWI using UGRs as anchors. Our data shows that a cut point of "4.5" can best differentiate between websites that users perceive as good vs. websites users perceive as bad. Furthermore, we were also able to show that this cut point was relatively stable in a large number of pseudo samples. While this is no guarantee that the cut points will be replicated in other samples, it gives some confidence to use it for websites that were not part of the present set. In order to find converging evidence a second study was performed.

3. Study 2

In the second study, we tried to replicate and confirm the findings from the first study using users rankings of websites as anchors to determine optimal cut points for the VisAWI-S.

3.1. Method

3.1.1. Participants

A total of 354 participants took part in study 2; 179 were female (51 %). Ages ranged from 17 to 84 years (*M* = 46.82, *SD* = 14.64). The education level of 77 % of the participants was high school diploma or higher. On average, the participants had been using the Internet for 14.56 years (Min = 3, Max = 35, SD = 4.67) and stated an active use of on average 2.4 hours a day (Min = .3, Max = 14, SD = 2.01). Participants took part voluntarily and on an anonymous basis. They got no compensation but were able to take part in a lottery of book vouchers at the end of the study.

3.1.2. Stimulus material

Screenshots of ten websites, different from the ones in study one, were chosen from ten different web content domains (like done in the previous study). To minimize the effect of other variables like content these websites were translated in Finnish language with the help of Google Translate, Images were edited with Gimp 2.8. Thus, stimuli were incomprehensible for German participants but due to the similarity of Finnish and German in typography still comparable in terms of the general design.

3.1.3. Procedure and Measures

Again, participants were recruited via the PsyWeb panel (https://psyweb.uni-muenster.de/). They received an invitation e-mail contained a link to the web-based study and were informed about aim and structure of the survey. After being asked for some demographic information (age, gender, education level, Internet experience), participants were presented with all websites from the stimulus set in random order. They were ask to rank the screenshots according to the visual appeal of the website. Afterwards, participants were asked in randomized order to rate each stimulus with the VisAWI-S (Moshagen and Thielsch 2013). Next, they were given the opportunity to comment on the study and, if wished, to exclude their data from the subsequent analysis. At the end of the survey, participants were thanked and had the opportunity to take part in a lottery of ten book vouchers with a value of ten Euro each (anonymity was guaranteed through the use of an individual winning code). On average, participants took 16 minutes to complete the study.

3.1.4. Data analysis

Data were analyzed in a similar fashion as before. First, we used scatter plots to describe the relation between the VisAWI-S scores and users' ranking of the website. Second, we dichotomized the ranks at the middle to yield a dichotomous rating ("good" vs. "bad" websites). Different cut points for the continuous VisAWI-S were evaluated as before using ROC analysis. Third, we use bootstrapping to quantify the variability of the findings.

3.2. Results

3.2.1. Correlations between rank and VisAWI-S ratings

Overall we found a large and highly significant correlation between the VisAWI-S and the rank (r = -.72; 95% CI = -.70 to -.74; p < .001) with higher VisAWI-S ratings indicating a better rank. As can be seen in figure 4 the mean VisAWI-S ratings in the different rank-categories can be fitted very well to a linear regression line.

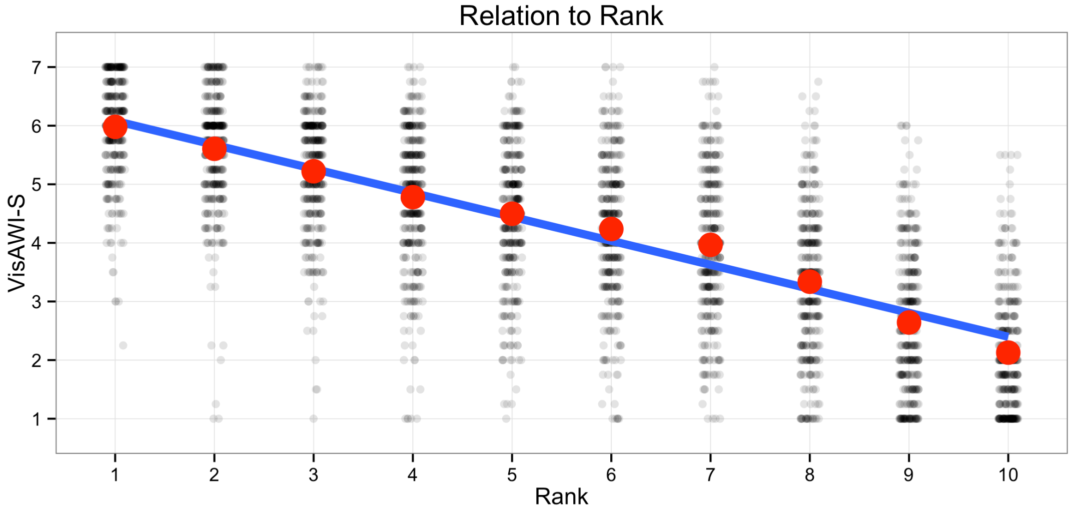

Figure 4. Relation between participants' ranking and VisAWI-S (Moshagen and Thielsch 2013) scores.
Note: Red dots indicate means. Error bars represent 95% CI for the mean. Straight lines represent best linear fit.

3.2.2. Optimal cut points for the whole sample

Inspection of the ROC-curve showed that the VisAWI-S scores could be used to distinguish between overall "good" vs. "bad" sites (fig. 5). The classification based on VisAWI-S yielded an AUC-value of .85 (95% CI .84 to .86).

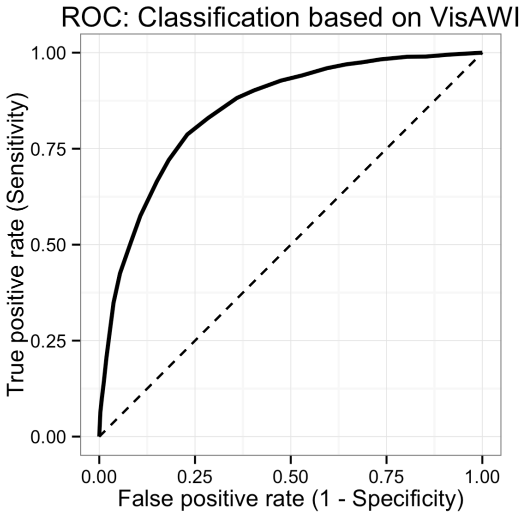

Figure 5. ROC-curve for the VisAWI-S (Moshagen and Thielsch 2013) against the dichotomous good vs. unattractive ranking.
Note: Broken line represents chance classification

The cutpoint that emerged as optimal in the full sample was "4.5" (sensitivity = .79; specificity = .77). As can be seen in table 2 several other cut points also yielded high levels of sensitivity and specificity.

Table 2. Performance of various alternative cut points in the full sample

cut point	sensitivity	specificity
7	0.06	1
6.75	0.1	0.99
6.5	0.14	0.99
6.25	0.21	0.98
6	0.35	0.96
5.75	0.43	0.95
5.5	0.5	0.92
5.25	0.58	0.89
5	0.66	0.85
4.75	0.72	0.82
4.5*	0.79	0.77
4.25	0.83	0.72
4	0.88	0.64
3.75	0.9	0.6
3.5	0.93	0.53
3.25	0.94	0.47
3	0.96	0.41
2.75	0.97	0.36
2.5	0.98	0.31
2.25	0.98	0.27
2	0.99	0.2
1.75	0.99	0.15
1.5	0.99	0.11
1.25	0.99	0.09
1	1	0

Note. * = Optimal cut point according to rank

3.2.3. Variability of optimal cut points

Our bootstrapping analysis showed that four different cut points were identified as optimal in some of the pseudo samples. The majority of bootstrapping samples (93.73%) yielded 4.5 as optimal cutpoint. Three other cut points 4.25, 4.75, and 4.00 emerged as optimal in 5.33%, .91%, and .03% of the pseudo samples.

3.3. Discussion

The aim of the second study was to use the same ROC-based method to define meaningful scores on the VisAWI-S with participants website rankings as anchor. Again we find that a cut point of “4.5” best differentiates between websites that users perceive as good vs. websites users perceive as bad. An inspection of the variability of these findings shows that this is relatively stable in the sense that this cut points was identified as optimal in the vast majority of pseudo samples. This further supports the contention that this is in fact a relevant cut point. Furthermore, we were also able to show that this cut point was relatively stable in a large number of pseudo samples.

4. General Discussion

The aim of the present studies was to define meaningful aesthetics scores on the VisAWI and to introduce the ROC-based method to define such scores for online user ratings on websites. In the following we will discuss the practical implications before turning to a methodological discussion.

4.1. Practical implications

An aesthetics score on the VisAWI higher than “4.5” is associated with an overall good impression of the website. Importantly we found this cutpoint consistently in two studies using different anchors. From a practical perspective this can be helpful from several points of view: First, it adds knowledge to the application and interpretation of the VisAWI. Before the present study an interpretation of general VisAWI sum scores was interpreted in terms of reported means (Moshagen and Thielsch 2010), now a much more precise cut point for interpretation is available. Second, our results indicate that improvements in aesthetics beyond this point may be of lesser importance than those up to “4.5”. To create a website perceived as rather good in terms of aesthetic a designer “just” need to pass this threshold of a “4.5”-evaluation, but has not the imperative to create a website reaching a 6-point or near 7-point evaluation on the VISAWI. Third, there might be occasions were you are not able to test your website against a benchmark with other websites – for example if the topic of your website is quite unique or you are entering a new market without any direct competitors who are already online. In this situation a cut point is again a good guideline to interpret website test data. Fifth, we hope that our example can inspire further research with other measures in HCI. In the same way a cut point is helpful in interpreting test data of website aesthetics, evaluations of usability, credibility, website content or other website measures can profit from this method.

4.2. Methods

On a methodological level we suggest ROC-based methods to define optimal cut points for HCI measures. Our main argument for ROC-based cut points is that they are more informative than alternative criteria such as mean plus two standard deviations (Copay et al. 2007). In the context of website evaluations this means that the latter do not give any information about users’ overall satisfaction with these sites. In contrast to this the sensitivity and specificity values that are associated cut points from ROC-based methods allow us to estimate how many users will be satisfied with a certain level of aesthetics.

While it is straightforward to select the cut point associated with the highest Youden index as optimal, three aspects of such an optimal cut point procedure are noteworthy:

1. The cut points critically depend on the anchor that is being used as a gold standard. Only if this anchor is both reliable and valid will the optimal cut points be beneficial to practitioners. For aesthetics the reliability of the gold standard might be improved by using paired-comparison data that are analyzed with multidimensional scaling techniques. While it is inherently hard to find external criteria for perceived aesthetics, it may be easier to find such anchors for constructs such as usability that are more closely related to observable behavior, e.g. whether or not participants are able to solve a specific task that involves an interface (cf. Bargas-Avila and Hornbæk 2011; Hornbæk 2006). If such a gold standard is not available statistical methods that detect and correct for the imperfect measurement of the gold standard (Erdfelder and Moshagen

2013; Reitsma et al. 2009) or triangulation using several gold-standards are the best ways to raise the level of confidence into specific cut points .

2. Once an optimal cut point has been determined in one specific sample, the associated sensititivity and specificity values in this sample are highly optimistic (Hirschfeld and do Brasil 2014; Leeflang et al. 2008). In other words optimal cut points that were determined for one specific sample, result in lower accuracies when they are applied to other samples. This problem can only be solved by prospectively evaluating pre-specified cut points.
3. The resulting optimal cut points are susceptible to chance variations within samples. That is, repeated studies using similar methods may yield different optimal cut points (Hirschfeld and do Brasil 2014). This problem can be addressed by using bootstrapping to estimate the random variability of the optimal cut points. Using such an estimate of the variability possible differences between samples and studies can be tested (Hirschfeld et al. 2014).

We believe that it is vital for the research community to invest more resources into the interpretation of existing measures and less into the development of novel measures. For practitioners the number of different questionnaires and scales for similar constructs in HCI might already be confusing. Adding more will not help answer practical questions, such as “is the website ready to launch?” or “do we need to invest more into the design?”. Future studies should try to systematically determine cut points for scales in HCI using methods to empirically define cut points that are meaningful to users and thus can be used by practitioners to inform everyday decisions.

4.3. Limitations and future research

There are several limitations that have to be kept in mind when interpreting the present findings. First, we have not investigated possible differences between subgroups of users and/or websites. We have recently shown that expert designers apply more stringent criteria when ratings the aesthetics of websites (Hirschfeld, Wachlin, and Thielsch 2013). While more research is necessary to investigate such differences in more detail this findings also highlights the need for tools that can be used by design experts to gather and interpret ratings of non-experts. Similarly it is possible that different cut points are necessary for various website domains, e.g. aesthetics may be a minor aspect for work-related websites while it may be of higher importance to leisure-related websites. However, as described above such comparisons need to take into account the random variability of optimal cut points (Hirschfeld et al. 2014). Second, all of the tested participants as well as the stimuli used shared the same cultural background. Some authors stress the importance of cultural factors in website design perception (Marcus and Gould 2000; Tractinsky 1997), especially for web design aspects like color and images (Cyr, Head, and Larios 2010). The extent to which our findings are prone to cultural differences could be analyzed by a cross-cultural approach. Third, given the myriad of existing websites, we tested only a limited sample of stimuli consisting of business and institutional websites. We excluded private websites from our study, as these show such a high variability in design, that it is hard to find a prototypical private website. Additional replication with different stimuli would be ideal to validate our results.

4.4. Summary and conclusion

While measures for website aesthetics have already been established, these are often only of limited use to practitioners who have to make decisions based on this data. We have shown that a score of “4.5” on the VisAWI is a meaningful cut point for users’ first and overall impression. On a more general level we hope that others find our approach useful in developing meaningful cut points for other measures in HCI such as usability, credibility or content perception.

References

Bargas-Avila, J.A., and Hornbæk, K. 2011. Old Wine in New Bottles or Novel Challenges: A Critical Analysis of Empirical Studies of User Experience. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, , 2689–98. ACM. http://dl.acm.org/citation.cfm?id=1979336.

Berlyne, D.E. 1971. *Aesthetics and Psychobiology*. East Norwalk, CT, US: Appleton-Century-Crofts.

———. 1974. *Studies in the New Experimental Aesthetics*. Washington D.C.: Hemisphere Publishing Corporation.

Bustamante, E.A., Bliss, J.P., and Anderson, B.L. 2007. Effects of Varying the Threshold of Alarm Systems and Workload on Human Performance. *Ergonomics*, 50 (7), 1127–47. doi:10.1080/00140130701237345.

Copay, A.G., Subach, B.R., Glassman, S.D., Polly, D.W., and Schuler, T.C. 2007. Understanding the Minimum Clinically Important Difference: A Review of Concepts and Methods. *The Spine Journal*, 7 (5), 541–46. doi:10.1016/j.spinee.2007.01.008.

Cyr, D., Head, M., and Larios, H. 2010. Colour Appeal in Website Design within and across Cultures: A Multi-Method Evaluation. *International Journal of Human-Computer Studies*, 68 (1), 1–21.

De Angeli, A., Sutcliffe, A., and Hartmann, J. 2006. Interaction, Usability and Aesthetics: What Influences Users’ Preferences? In *Proceedings of the 6th Conference on Designing Interactive Systems*, , 271–80. http://dl.acm.org/citation.cfm?id=1142446.

Erdfelder, E., and Moshagen, M. 2013. Conjoint Measurement of Disorder Prevalence, Test Sensitivity, and Test Specificity: Notes on Botella, Huang, and Suero’s Multinomial Model. *Frontiers in Psychology*, 4. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3837240/.

Fechner, G.T. 1860. *Elements of Psychophysics.* Leipzig: Breitkopf & Härtel. http://psycnet.apa.org/books/11304/026.

———. 1876. *Vorschule Der Ästhetik*. Leipzig: Breitkopf & Härtel.

Flavián, C., Guinalíu, M., and Gurrea, R. 2006. The Role Played by Perceived Usability, Satisfaction and Consumer Trust on Website Loyalty. *Information & Management*, 43 (1), 1–14.

Fluss, R., Faraggi, D., and Reiser, B. 2005. Estimation of the Youden Index and Its Associated Cutoff Point. *Biometrical Journal*, 47 (4), 458–72. doi:10.1002/bimj.200410135.

Hirschfeld, G., and do Brasil, P.E.A.A. 2014. A Simulation Study into the Performance of “optimal” Diagnostic Thresholds in the population:“Large”

Effect Sizes Are Not Enough. *Journal of Clinical Epidemiology*, 67 (4), 449–53.

Hirschfeld, G., Wachlin, L., and Thielsch, M.T. 2013. Was macht Webdesign-Experten aus? Eine Signalentdeckungs-Analyse. In S. Boll, S. Maaß & R. Malaka (Eds.), Mensch & Computer 2013 (S. 273-276). München: Oldenbourg.

Hirschfeld, G., Wager, J., Schmidt, P., and Zernikow, B. 2014. Minimally Clinically Significant Differences for Adolescents With Chronic Pain—Variability of ROC-Based Cut Points. *The Journal of Pain*, 15 (1), 32–39.

Hornbæk, K. 2006. Current Practice in Measuring Usability: Challenges to Usability Studies and Research. *International Journal of Human-Computer Studies*, 64 (2), 79–102.

Jacobsen, T. 2006. Bridging the Arts and Sciences: A Framework for the Psychology of Aesthetics. *Leonardo*, 39 (2), 155–62.

Lavie, T., and Tractinsky, N. 2004. Assessing Dimensions of Perceived Visual Aesthetics of Web Sites. *International Journal of Human-Computer Studies*, 60 (3), 269–98.

Lee, S., and Koubek, R.J. 2012. Users' Perceptions of Usability and Aesthetics as Criteria of Pre–and Post–use Preferences. *European Journal of Industrial Engineering*, 6 (1), 87–117.

Leeflang, M.M.G., Moons, K.G.M., Reitsma, J.B., and Zwinderman, A.H. 2008. Bias in Sensitivity and Specificity Caused by Data-Driven Selection of Optimal Cutoff Values: Mechanisms, Magnitude, and Solutions. *Clinical Chemistry*, 54 (4), 729–37.

Marcus, A., and Gould, E.W. 2000. Crosscurrents: Cultural Dimensions and Global Web User-Interface Design. *Interactions*, 7 (4), 32–46.

Moshagen, M., Musch, J., and Göritz, A.S. 2009. A Blessing, Not a Curse: Experimental Evidence for Beneficial Effects of Visual Aesthetics on Performance. *Ergonomics*, 52 (10), 1311–20.

Moshagen, M., and Thielsch, M.T. 2010. Facets of Visual Aesthetics. *International Journal of Human-Computer Studies*, 68 (10), 689–709.

———. 2013. A Short Version of the Visual Aesthetics of Websites Inventory. *Behaviour & Information Technology*, 32 (12), 1305–11. doi:10.1080/0144929X.2012.694910.

Reitsma, J.B., Rutjes, A.W.S., Khan, K.S., Coomarasamy, A., and Bossuyt, P.M. 2009. A Review of Solutions for Diagnostic Accuracy Studies with an Imperfect or Missing Reference Standard. *Journal of Clinical Epidemiology*, 62 (8), 797–806. doi:10.1016/j.jclinepi.2009.02.005.

Schisterman, E.F., and Perkins, N. 2007. Confidence Intervals for the Youden Index and Corresponding Optimal Cut-Point. *Communications in Statistics—Simulation and Computation®*, 36 (3), 549–63.

Sonderegger, A., and Sauer, J. 2010. The Influence of Design Aesthetics in Usability Testing: Effects on User Performance and Perceived Usability. *Applied Ergonomics*, 41 (3), 403–10.

Swets, J.A. 1988. Measuring the Accuracy of Diagnostic Systems. *Science*, 240 (4857), 1285–93.

Thielsch, M.T., Blotenberg, I., and Jaron, R. 2014. User Evaluation of Websites: From First Impression to Recommendation. *Interacting with Computers*, 26 (1), 89–102. doi:10.1093/iwc/iwt033.

Thielsch, M.T., and Hirschfeld, G. 2010. High and Low Spatial Frequencies in Website Evaluations. *Ergonomics*, 53 (8), 972–78. doi:10.1080/00140139.2010.489970.

———. 2012. Spatial Frequencies in Aesthetic Website Evaluations--Explaining How Ultra-Rapid Evaluations Are Formed. *Ergonomics*, 55 (7), 731–42. doi:10.1080/00140139.2012.665496.

Thielsch, M.T. and Moshagen, M. 2011. Erfassung visueller Ästhetik mit dem VisAWI. In H. Brau, A. Lehmann, K. Petrovic & M. C. Schroeder (Eds.), Usability Professionals 2011 (S. 260-265). Stuttgart: German UPA e.V..

Tractinsky, N. 1997. Aesthetics and Apparent Usability: Empirically Assessing Cultural and Methodological Issues. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, , 115–22. http://dl.acm.org/citation.cfm?id=258626.

Tuch, A.N., Roth, S.P., Hornbæk, K., Opwis, K., and Bargas-Avila, J.A. 2012. Is Beautiful Really Usable? Toward Understanding the Relation between Usability, Aesthetics, and Affect in HCI. *Computers in Human Behavior*, 28 (5), 1596–1607.