How Informative is Informative? Benchmarks and Optimal Cut Points for E-Health Websites

Meinald T. Thielsch
Department of Psychology
University of Münster
Münster, Germany
thielsch@uni-muenster.de

Carolin Thielsch
Department of Psychology
University of Münster
Münster, Germany
carolinthielsch@uni-muenster.de

Gerrit Hirschfeld
Faculty of Business and Health
Bielefeld University of Applied Sciences
Bielefeld, Germany
gerrit.hirschfeld@fh-bielefeld.de

ABSTRACT

Scores of different evaluation measures resulting from website tests are difficult to interpret without comparative data. Benchmarks and optimal cut points provide such interpretation aids. Benchmarks are usually built with test score means based on a tested pool of comparable websites. Optimal cut points are calculated with an external criterion using receiver-operating-characteristic (ROC) based methods applied on website evaluations. Due to relevance and sensitivity of the topic, making the right decision based on evaluation data is of particular importance for creators and owners of websites presenting health-related information. Thus, we combined data of two studies, with a total of n = 2.614 participants, evaluating m=33 health-related websites. Established questionnaires were applied: Web-CLIC (website content), PWU-G and UMUX-Lite (usability), VisAWI-S (aesthetics), and trusting belief scales of McKnight et al. [7]. We calculated overall and specific values for four categories of e-health websites. Benchmarks were quite comparable among categories while optimal cut points differed more. Particularly, cut points were high for charity websites and partly lower for the category "Personal sites & support groups". In general, user requirements for e-health websites appear to be significantly higher than available published benchmarks and cut points for websites in other areas.

CCS CONCEPTS

• Human-centered computing → Human computer interaction (HCI) → Empirical studies in HCI

MuC'19 Workshops, Hamburg, Deutschland

https://doi.org/10.18420/muc2019-ws-642

KEYWORDS

Benchmarks, Cut points, Evaluation, E-Health, Health websites

ACM Reference format:

Meinald T. Thielsch, Carolin Thielsch, and Gerrit Hirschfeld. 2019. How informative is informative? Benchmarks and optimal cut points for E-Health Websites. In *Mensch und Computer 2019 – Workshopband*, Bonn: Gesellschaft für Informatik e.V., https://doi.org/10.18420/muc2019-ws-642

1 Introduction

In Germany, more than 90% of people over the age of 14 are online and about three quarters of the population use the internet daily [3]. When searching for information, users very quickly and spontaneously make a selection from the large number of available websites [13]. The subjective experience of a website – the user experience – is central to this decision. Content, usability, and aesthetics perceptions as well as trust towards a website and its provider are central dimensions of web users' experiences (e.g., [2], [10], [13], [15]). Each dimension is crucial for users' acceptance, appreciation, revisit and recommendation of specific websites. For example, if users do not understand the content or distrust the website provider, they will reorient themselves and search for another website. Thus, it is necessary to evaluate websites to understand the perception and impressions of its users.

1.1 Website Evaluation

In general, evaluations in practice should support decision-making processes - and deliver results that are as reliable and as valid as possible. Evaluations help to describe and evaluate actual conditions. Only when the situation at hand has been accurately described will it become clear whether and which further measures are necessary. If this description is incorrect, there is a danger that interventions will be based on incorrect assumptions or that projects will fail because necessary steps will not be identified and implemented.

Website evaluation is primarily dedicated to the perception and behavior of website visitors. Here, it is very helpful to distinguish between different phases in which users react to different aspects of a website. Initially, the first impression of a website is mainly determined by its aesthetics. Visual aesthetics is very quickly perceived and evaluated in the first hundred milliseconds of use [1]. The evaluation of the content requires reflected cognitive processes and therefore probably takes a little longer [13]. First impressions regarding the credibility of a website can be given after about three to four seconds [10]. In order to evaluate the usability of a website in a meaningful way, a real interaction with it and thus additional time is necessary (e.g., [4], [16]).

The need for highly reliable and valid instruments in website evaluation is obvious. Yet, another import factor is the interpretation of evaluation results: All widely-used evaluation tools yield continuous scores, for example leading to a website usability score of 5.5. Within themselves these scores are difficult to interpret. One possibility is to compare a given website evaluation with ratings of prior versions of the website or with another comparable website in the field (A/B testing). As this requires extra work and additional resources, benchmarks and cut points are a valuable alternative. Both enable a meaningful interpretation of individual test scores supporting the decision process following an evaluation. Due to the far-reaching consequences of the topic, this is of particular importance for websites presenting health-related information, particularly as two in three German Internet users search for health information online [9].

1.2 Benchmarks and Optimal Cut Points

Benchmarks and cut points are interpretation aids supporting the decision process when assessing website evaluation data. Benchmarks are usually based on a pool of comparable websites tested. They could be presented in form of mean and standard deviations of summed previous website ratings. Thus, benchmarks enable a comparison to a potentially large pool of other sites. This allows to assess, for example, whether a specific website is perceived as more or less informative than an average site from a given test pool.

Yet, the creation of a benchmark pool can be time-consuming and resource-intensive. In consequence, they are provided only by few questionnaire authors in HCI. Furthermore, benchmarks do not offer information on the relevance of specific values: For example, even if the content of a specific website receives above-average ratings, that does not necessarily imply that users are satisfied with the presented website.

To solve such problems, optimal cut points can serve as an orientation. They consist of critical values that indicate, for example, when a user will classify a website as generally good or bad. Optimal cut points are determined using receiver-operating-characteristic (ROC) based methods (see [5]) applied on website evaluations. This procedure is inspired by methods in medicine and needs an external criterion, for example global ratings of the overall impression of website users or a ranking of websites [5]. In contrast to benchmarks, they require a substantial but comparatively smaller sample size, especially if there are large differences between positive and negative stimuli.

The aim of the present study is to provide aids for interpreting individual evaluation scores of several instruments when assessing e-health websites by providing both, benchmarks and optimal cut points.

2 Methods

We combined data of two studies, with a total of n = 2.614 participants, evaluating m = 33 health-related websites. Study 1 (n = 349, 48% female, Mage = 47.81 years, aged 18 – 82, m = 3) used a within-subject design; Study 2 (n = 2265, 49% female, Mage = 51.80 years, aged 16 – 79, m = 30) used a between-subject design. Study 1 participants were recruited via the panel PsyWeb (https://psyweb.uni-muenster.de/), Study 2 participants via a commercial panel.

The website pool was based on evaluations of seven experts (including one of the studies' authors). The 33 tested websites were clustered in four different categories:

1. Government & educational establishment websites (m = 10 websites)
2. Commercial health news and information websites (m = 12 websites)
3. Charity sites (m = 5 websites)
4. Personal sites & support groups (m = 4 websites)

At the beginning of both studies, the participants were informed about objectives, involved researchers, anonymity, voluntariness and duration. If they agreed to the terms, they could start the survey and first provide demographic information. Then, in Study 1, three websites were randomly presented to all participants; in Study 2 participants were randomly assigned to one website from the stimulus set. Each website was presented fully-functional and rated by 60 to 349 participants (Mean = 105). In both studies, participants were given the task to freely explore the website, including the use of subpages and without any time pressure (the execution of the task was controlled by a Java script.). Afterwards, they were asked to evaluate the given website using a batterie of established and validated web site questionnaires. Content perceptions were gathered with the Web-CLIC [13]. Usability was judged with the aid of two measures: The PWU-G scale (original: [2]; German version: [11], [16]) and the UMUX-Lite [6]. Visual aesthetics was evaluated with the short version of the Visual Aesthetics of Websites Inventory (VisAWI-S, [8]). Trust evaluations were gathered applying the trusting belief scales of McKnight, Choudhury and Kacmar [7]. All items were scaled on seven-point Likert scales ranging from 1 ("strongly disagree") to 7 ("strongly agree").

Additionally, the overall website impression was assessed with a grade on a on a six-point grading scale ("What overall rating do you give this website?", 1 = "very good", 2 = "good", 3 = "satisfactory", 4 = "adequate", 5 = "poor", 6 = "unsatisfactory")

commonly used in the German education system. Both studies assessed further variables not pertinent to the present analyses. At the end, participants could exclude their data from the subsequent analysis and were thanked for their participation.

3 Results

First, we analyzed whether the formed four categories of e-health websites were distinct. Based on prior research, we included additional three covariates (age, gender, level of education) to control for potential biases. We found significant differences between the four categories in a MANCOVA ($F (30, 9477) = 13.894$, $p < .01$, $\eta^2 = .042$). Covariates showed lower yet significant effect sizes: age ($F (10, 3157) = 10.455$, $p < .01$, $\eta^2 = .032$), gender ($F (10, 3157) = 4.815$, $p < .01$, $\eta^2 = .015$) and level of education ($F (10, 3157) = 5.323$, $p < .01$, $\eta^2 = .017$). In the univariate tests significant differences were revealed for all instruments except the UMUX-Lite.

3.1 Benchmarks

Based on the relatively smaller effect sizes of covariates, we decide to calculate benchmark values for the four categories and an overall benchmark, but not to standardize values psychometrically for different age groups, gender or level of education. As indicated by the effect size in the MANCOVA, mean differences for the website categories were mostly rather small. However, two systematic differences are particularly striking: First, benchmarks are lower for the category "Commercial health news and information websites" when it comes to content credibility and trust measures. Second, benchmarks are lower for "Charity sites" regarding perceived visual aesthetics. All benchmarks can be found in Table 1.

3.2 Optimal Cut Points

Optimal cut points were determined based on users` overall impression measured with the grading scale as an external criterion. We used the Youden-index to identify the cut point that best differentiated between bad (grade 3 and below) and good (grade 2 and higher) websites (as done by [5]). We observed more differences in the optimal cut point values between the four categories than before in the benchmark analysis. Particularly, the cut points for "Charity sites" were high for several scales (overall WEB-CLIC score, informativeness, credibility, UMUX-LITE, integrity, and competence). The three lowest cut points were found for the category "Personal sites & support groups" (for overall WEB-CLIC score, likeability, and competence). All cut points can be found in Table 2.

Table 1: Benchmark data in form of means for each website category as well as overall means

	Government & educational establishment websites (n = 740, m = 10)	Commercial health news and information websites (n = 937, m = 12)	Charity sites (n = 512; m = 5)	Personal sites & support groups (n = 286, m = 4)	Overall (n = 2.614, m = 33)
Content					
Overall score (Web-CLIC mean)	5.22 (1.05)	5.09 (1.11)	5.16 (1.00)	5.06 (1.03)	5.15 (1.05)
Clarity (Web-CLIC)	5.52 (1.05)	5.44 (1.03)	5.29 (1.07)	5.46 (1.05)	5.41 (1.05)
Likeability (Web-CLIC)	4.64 (1.35)	4.64 (1.40)	4.59 (1.35)	4.26 (1.39)	4.58 (1.37)
Informativeness (Web-CLIC)	5.40 (1.11)	5.26 (1.16)	5.38 (1.06)	5.32 (1.07)	5.35 (1.10)
Credibility (Web-CLIC)	5.32 (1.13)	5.04 (1.17)	5.39 (1.02)	5.22 (1.05)	5.26 (1.10)
Usability					
Usability (PWU-G)	5.64 (1.11)	5.55 (1.07)	5.37 (1.23)	5.61 (1.14)	5.51 (1.15)
Usability (UMUX-Lite)	5.59 (1.12)	5.57 (1.09)	5.44 (1.28)	5.60 (1.12)	5.54 (1.18)
Aesthetics					
Aesthetics (VisAWI-S)	5.26 (1.21)	5.20 (1.17)	4.72 (1.42)	5.15 (1.27)	5.04 (1.31)
Trust					
Benevolence	4.88 (1.12)	4.49 (1.24)	4.81 (1.10)	4.94 (1.18)	4.75 (1.17)
Integrity	4.98 (1.06)	4.66 (1.12)	4.96 (1.02)	5.01 (1.09)	4.89 (1.08)
Competence	5.23 (1.10)	4.97 (1.20)	5.16 (1.12)	5.22 (1.08)	5.14 (1.14)

Note. All measures were scaled on seven-point Likert scales ranging from 1 ("strongly disagree") to 7 ("strongly agree").

Table 2: Optimal cut points for each website category and overall cut points

	Government & educational establishment websites (n = 740, m = 10)	Commercial health news and information websites (n = 937, m = 12)	Charity sites (n = 512; m = 5)	Personal sites & support groups (n = 286, m = 4)	Overall (n = 2.614, m = 33)
Content					
Overall score (Web-CLIC mean)	5.00	5.00	5.33	4.67	4.83
Clarity (Web-CLIC)	5.67	5.33	5.67	5.33	5.67
Likeability (Web-CLIC)	4.33	4.67	4.67	3.67	4.67
Informativeness (Web-CLIC)	5.33	5.33	5.67	5.00	5.33
Credibility (Web-CLIC)	5.00	5.00	5.67	5.33	5.33
Usability					
Usability (PWU-G)	5.14	5.43	5.29	5.57	5.29
Usability (UMUX-Lite)	5.50	5.50	6.00	5.50	5.50
Aesthetics					
Aesthetics (VisAWI-S)	5.25	5.00	4.75	5.00	5.00
Trust					
Benevolence	4.67	4.33	4.33	4.33	4.33
Integrity	4.75	4.50	5.25	4.75	4.50
Competence	5.00	5.00	5.25	4.50	5.00

Note. All measures were scaled on seven-point Likert scales ranging from 1 ("strongly disagree") to 7 ("strongly agree").

4 Discussion

The present analyses provided benchmarks and optimal cut points for several evaluation instruments for the domain of health-related websites. Thus, comparisons with existing websites and estimates for the evaluation of new websites are provided for this type of websites. Found benchmark and cut point values were mostly similar for the analyzed instruments. Both indicated, for example, that aesthetic values above 5.0 on the VisAWI-S are desirable. Yet, there are some differences, for example is the UMUX-Lite benchmark for charity websites 5.4, the optimal cut point is 6.0. When taking a closer look to the results of the present analyses three aspects stand out:

1. Benchmark values are mostly quite similar among the different categories of e-health websites. For most situations it seems therefore possible to simply use the general benchmarks instead of the more specific ones.
2. Optimal cut points seem to differ a little more than benchmarks. Therefore, depending on the situation, the specific cut point values of a website category might be taken into account.
3. Compared to available benchmarks ([12], [14], [14]) and published cut points ([5], [13]) user demands on e-health websites seem to be significantly higher, since all values found in the present analyses are (in some cases very clearly) higher. This may be due to the importance that website users attach to health topics.

Some limitations should be taken into account when interpreting the results of the present analyses. At the same time, however, they offer opportunities for future research: First, given the enormous number of e-health websites and its users, the present data cannot be seen as fully representative. In particular, the number of tested websites in categories 3 (Charity sites) and 4 (Personal sites & support groups) is rather low. Therefore, further studies on health websites are worthwhile. Second, all study participants originated from Germany and thus shared a common cultural background. However, the perception of health information online may be different in other cultures. Third, due to time limitations, we tested only with a part of the available validated website evaluation tools. There are several other tools that could be investigated in the same manner in future. Fourth, for the optimal cut point analyses, one might argue that there are better external criteria as an overall grade (see the discussion provided in [5]). Thus, replications of our findings and further analyses of potential interpretation aids for e-health website evaluations are highly welcome.

From a practical point of view, it might it be much easier to work towards higher values than the comparisons presented here than to try to reach the top of each scale. In any case, benchmarks and cut points allow a meaningful interpretation of test scores beyond the original pure numerical value. Thus, we hope that both, scientists evaluating existing e-health services as well as practitioners creating new ones, find the provided interpretation aids as useful.

ACKNOWLEDGMENTS

The authors would like to express their sincere thanks to the experts involved for their support. In addition, we are grateful for the award of a grant supporting the sampling of the two studies: This research is supported by the Federal Centre for Health Education (BZgA) on behalf of the Federal Ministry of Health.

REFERENCES

[1] J. Bölte, T. Hösker, G. Hirschfeld, & M. T. Thielsch (2017). Electrophysiological correlates of aesthetic processing of webpages: A comparison of experts and laypersons. PeerJ, 5, e3440. https://doi.org/10.7717/peerj.3440

[2] C. Flavián, M. Guinalíu, & R. Gurrea (2006). The role played by perceived usability, satisfaction and consumer trust on website loyalty. Information & Management, 43(1), 1-14. http://doi.org/10.1016/j.im.2005.01.002

[3] V. B. Frees, & W. Koch (2018). ARD / ZDF-Onlinestudie 2018: Zuwachs bei medialer Internetnutzung und Kommunikation. Media Perspektiven, (9/2018), 398–413.

[4] K. Hornbæk (2006). Current practice in measuring usability: Challenges to usability studies and research. International Journal of Human-Computer Studies, 64(2), 79–102. https://doi.org/10.1016/j.ijhcs.2005.06.002

[5] G. Hirschfeld, & M. T. Thielsch (2015). Establishing meaningful cut points for online user ratings. Ergonomics, 58(2), 310–320. https://doi.org/10.1080/00140139.2014.965228

[6] J. R. Lewis, B. S. Utesch, & D. E. Maher (2013). UMUX-Lite: when there's no time for the SUS. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2099-2102). ACM. http://doi.org/10.1145/2470654.2481287

[7] D. H. McKnight, V. Choudhury, & C. Kacmar (2002). Developing and validating trust measures for e-commerce: An integrative typology. Information systems research, 13(3), 334-359. http://doi.org/10.1287/isre.13.3.334.81

[8] M. Moshagen, & M. T. Thielsch (2013). A short version of the visual aesthetics of websites inventory. Behavior & Information Technology, 32(12), 1305-1311. https://doi.org/10.1080/0144929X.2012.694910

[9] L. Nölke, M. Mensing, A. Krämer, & C. Hornberg (2015). Sociodemographic and health-(care-)related characteristics of online health information seekers: A cross-sectional German study. BMC Public Health, 15(1), 1–12. https://doi.org/10.1186/s12889-015-1423-0

[10] D. Robins, & J. Holmes (2008). Aesthetics and credibility in web site design. Information Processing & Management, 44(1), 386–399. https://doi.org/10.1016/j.ipm.2007.02.003

[11] M. T. Thielsch (2008). *Ästhetik von Websites. Wahrnehmung von Ästhetik und deren Beziehung zu Inhalt, Usability und Persönlichkeitsmerkmalen*. Münster: MV Wissenschaft.

[12] M. T. Thielsch (unter Mitarbeit von M. Salaschek) (2017). *Toolbox zur kontinuierlichen Website-Evaluation und Qualitätssicherung (Version 2.0)*. Arbeitsbericht, Köln: Bundeszentrale für gesundheitliche Aufklärung (BZgA). http://doi.org/10.17623/BZGA:224-2.0

[13] M. T. Thielsch, & G. Hirschfeld (2019). Facets of website content. Human-Computer Interaction, 34 (4), 279-327. http://doi.org/10.1080/07370024.2017.1421954

[14] M. T. Thielsch, & M. Moshagen (2015). *VisAWI Manual (Visual Aesthetics of Websites Inventory) and the short form VisAWI-S (Short Visual Aesthetics of Websites Inventory)*. https://doi.org/10.13140/RG.2.1.3985.6169

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

[16] M. T. Thielsch, R. Engel, & G. Hirschfeld (2015). Expected usability is not a valid indicator of experienced usability. PeerJ Computer Science, 1, e19. https://doi.org/10.7717/peerj-cs.19