Quick assessment of web content perceptions

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

[1]*Department of Psychology, University of Münster, Germany,*

[2]*Faculty of Business and Health, University of Applied Sciences Bielefeld, Germany*

Address for correspondence: Meinald T. Thielsch, University of Münster, Department of Psychology, Fliednerstr. 21, 48149 Münster, Germany | E-Mail: thielsch@uni-muenster.de | ORCID: 0000-0001-8493-9071

Short biographical notes on all contributors

Meinald T. Thielsch (thielsch@uni-muenster.de, www.meinald.de) is a psychologist with an interest in human-computer interaction and user experience; he is a tenured faculty member and an extraordinary professor of Organizational Psychology and Human-Computer Interaction in the Department of Psychology, University of Münster.

Gerrit Hirschfeld (gerrit.hirschfeld@fh-bielefeld.de, www.gerrithirschfeld.de) is a psychologist with an interest in diagnostics and research methods; he is a professor of Research Methods, Diagnostics, and Applied Psychology in the Faculty of Business and Health, University of Applied Sciences Bielefeld.

This is an Accepted Manuscript of an article published by Taylor & Francis in International Journal of Human-Computer Interaction on 19/08/2020, available online: https://www.tandfonline.com/doi/full/10.1080/10447318.2020.1805877.

Citation:
Thielsch, M. T. & Hirschfeld, G. (in press). Quick assessment of web content perceptions. *International Journal of Human-Computer Interaction.* https://doi.org/10.1080/10447318.2020.1805877

Quick assessment of web content perceptions

ABSTRACT

In digital media and on the World Wide Web, content is king. As such, users' subjective perceptions of content can influence a variety of their evaluations, thereby altering their attitudes and behavioral outcomes. Thus, users' content perceptions need to be assessed using a valid measure, but this often has to be done while keeping the survey time as short as possible. For these situations, we created a four-item short version of the Web-CLIC questionnaire (Thielsch & Hirschfeld, 2019). We tested this version, called the Web-CLIC-S, in a series of three studies, including 1,414 participants and 33 fully functional websites of different content domains. Confirmatory factor analysis confirms that the Web-CLIC-S reflects an unidimensional g-factor of subjective web content. The Web-CLIC-S also demonstrates high internal consistencies and high short- to medium-term retest reliabilities. Furthermore, we find strong evidence for construct validity in terms of convergent, divergent, discriminative, concurrent, incremental, and predictive validity. In a fourth study, encompassing 12,568 ratings on 183 websites, we provide benchmarks for 12 different content domains and optimal cut points. Overall, the present research suggests that the Web-CLIC-S can serve as a sound screening tool to assess users' subjective perception of content in research and practice settings.

Keywords: Evaluation; website content; information quality; user experience; Web-CLIC-S

Data availability: The data of the studies (including codebooks) are available at https://doi.org/10.5281/zenodo.3813293

1. INTRODUCTION

In modern digital media, content is king, as website users consider content to be the most important element for evaluation (Thielsch, Blotenberg & Jaron, 2014). In the context of the fast-paced Internet, organizations face enormous competition for online visitors, but the attention span of online users is extremely short: Decisions about whether to visit a website are made quickly, and the average length of stay is often little more than a minute (Liu, White, & Dumais, 2010). Compared to classic print media, readers of online material are hardly bound to one source and can switch to other related websites quickly using powerful search engines. Therefore, a website's content must be optimized with regard to both search engines and the actual users, who quickly determine whether to stay or go.

Thus, to optimize their web content, organizations need to know how users perceive the presented information, namely whether they understand, believe, and appreciate the content. Yet, most available evaluation tools only assess general website quality or only test website content with unidimensional single items and unaudited ad hoc scales (see Thielsch & Hirschfeld, 2019). An exception is the Web-CLIC questionnaire (Thielsch & Hirschfeld, 2019), which has been extensively validated and assesses four dimensions of website content. This is a sound instrument for a detailed and comprehensive evaluation, but it is less suitable for situations where the survey time must be kept to an absolute minimum and where only an overall evaluation of web content is necessary rather than a detailed assessment of multiple facets of content.

Such "short and sweet" evaluation situations are especially relevant when websites must be continuously monitored or when one wants to simultaneously evaluate a whole series of constructs. The amount of questions is a highly critical issue for many user and customer surveys in applied contexts. Moreover, in many research contexts, scientists scrutinize various other website aspects but still might want to assess how users are perceiving the content in general. For example, studies that focus on the effectiveness of online health interventions would benefit from concise measures of content, particularly as content perceptions can be influenced by users' health status (e.g., Bansal, Zahedi &

Gefen 2010; Thielsch & Thielsch, 2018). Additionally, there are situations were just a short manipulation check is needed, for example when a researcher manipulates other website aspects but wants to ensure that the content perception was not affected. Thus, the aim of the present work was to develop and validate a short version of the Web-CLIC for these various purposes.

1.1. Related work

Definition of website content

From a technical point of view, ISO 9241-151 defines content as "a set of content objects" and defines a content object as an "interactive or non-interactive object containing information represented by text, image, video, sound or other types of media" (ISO, 2006, p. 3). In the present work, we focus on subjective perceptions of web content that can be assessed using a survey approach and can be rated by typical users. Based on current theories of how users process websites (see Moshagen and Thielsch, 2010; Thielsch & Hirschfeld, 2019), we follow an interactionist perspective: The formation of subjective perceptions relies on the interaction between the characteristics of the perceiver, the use scenario, and the properties of web content objects (as defined in ISO 9241-151; ISO, 2006). Thus, we define the subjective part of website content as users' general perceptions, impressions, and ratings that result from their interaction with the presented content objects of a website (cf. Thielsch & Hirschfeld, 2019).

Importance of content perceptions

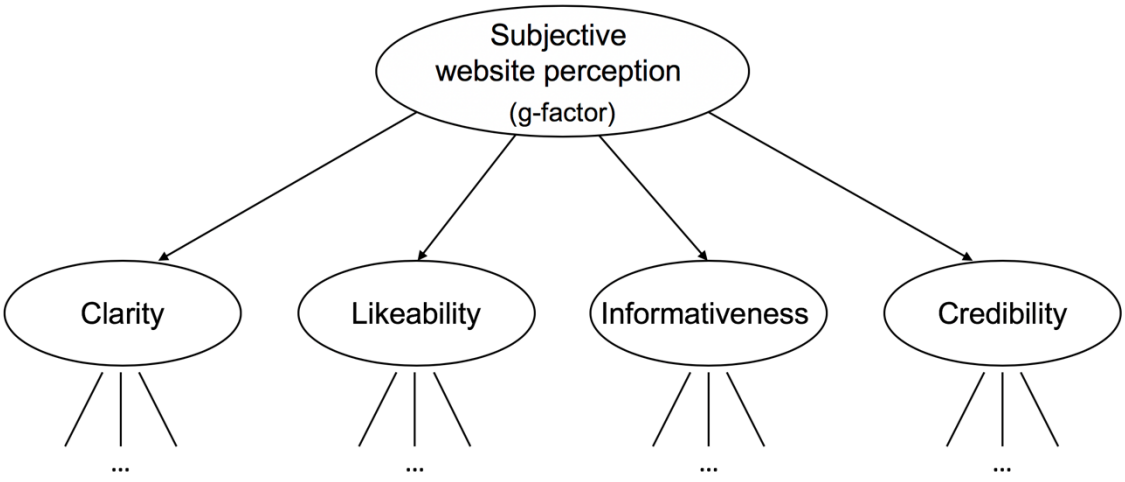
Users' perceptions of content correlate with other central constructs of user experience, such as overall attitudes and satisfaction (e.g., Kang & Kim, 2006; Shukla, Sharma, & Swami, 2010), perceived ease of use, usefulness and usability (e.g., Ahn, Ryu, & Han, 2007; Thielsch et al., 2014), and aesthetics (e.g., Moshagen & Thielsch, 2010; Robins & Holmes, 2008). Moreover, users' ratings of content quality are highly related to website success, users' loyalty, and users' intentions to revisit or recommend it (e.g., Kim & Niehm, 2009; Shukla et al., 2010). Further, users' ratings of content, above and beyond

global ratings, have also been found to predict their behavioral choices (Thielsch & Hirschfeld, 2019). Finally, content perceptions can influence task performance: Meeßen and colleagues (2020) systematically varied the credibility of information provided by a digital managing information system, and higher credibility led users to have greater trust in the system, resulting in better user performance and better user well-being. This illustrates that users' subjective perceptions of content are important and can affect actual behavior and related outcomes.

Assessing subjective perceptions of web content

In research, web users' content perceptions have been examined using five different survey strategies: (1) attribute lists and checklists, (2) single-item assessments (partly enclosed in general website evaluation scales), (3) unidimensional scales, (4) multidimensional scales enclosed in extensive measures of "website quality" and (5) specific instruments designed to assess perceptions of website content. Strategy (1) is well suited to expert assessments, but it is difficult to use to evaluate users' perceptions of content aspects. Further, most of the existing tools in strategies (2) to (5) suffer from shortcomings in reliability and validity (for a detailed review, see Thielsch & Hirschfeld, 2019). An exception is the Web-CLIC (Thielsch & Hirschfeld, 2019), which stands for **Web**site - **C**larity, **L**ikeability, **I**nformativeness, and **C**redibility. These four content areas are covered by 12 questions, and an overall sum score can be calculated, which reflects the user's subjective experience of the website content as a whole in the form of a g-factor (see Figure 1).

Figure 1: Structural model of the Web-CLIC (Thielsch & Hirschfeld, 2019)



Construction and validation of the Web CLIC were based on six studies with a total of n = 3,106 respondents and m = 60 tested websites (Thielsch & Hirschfeld, 2019). The questionnaire has proven to be reliable: The internal consistency (Cronbachs α) was at least >.80 for the scales and >.90 for the sum score representing the g-factor. The time stability (retest reliability) over a period of two weeks was in the range of $.69 \leq r \leq .81$ for the scales and $r = .84$ for the sum score. Regarding validity, comprehensive empirical evidence has been found through tests on factorial, convergent, divergent, discriminatory, competitive, experimental and predictive validity. For example, the authors found that the Web-CLIC specifically and accurately recorded experimental variations in the credibility of a health information website and was also able to predict the respondents' donation behavior for non-profit organizations. As further interpretation aids, optimal cut points and benchmarks for ten different website categories have been identified, based on 7,379 ratings of m = 120 websites (see Study 7 in Thielsch & Hirschfeld, 2019).

1.2. Aims of the present research

The aim of the present research was to create, validate, and benchmark a short version of the Web-CLIC (called Web-CLIC-S). In Study 1, we identified the most relevant items for evaluating web content from the original questionnaire. We then tested this newly created instrument with confirmatory factor analysis and analyzed it for reliability,

specifically in terms of internal consistency and test-retest reliability. Study 2 focused on the construct validation of the Web-CLIC-S and applied convergent, divergent, concurrent, and discriminative validation strategies. In Study 3, we further analyzed the validity and, particularly, the usefulness of the new measure; the usefulness was analyzed by 1) comparing results with global ratings and 2) predicting behavioral intentions and actual behavior.[1] In addition to these thorough psychometric analyses, in Study 4 we give advice for interpretation and practical use by providing benchmarks as well as optimal cut points.

2. STUDY 1: RELIABILITY AND FACTOR STRUTURE

The aim of Study 1 was to identify items that capture the essence of website content perceptions as measured with the full Web-CLIC. In addition, as it is essential to test the internal consistency, retest-reliability and unidimensionality of the shortened questionnaire, we therefore conducted a longitudinal study with three data collection points and time gaps of two days to two weeks between collection points.

2.1. Method

Participants

A total of 764 participants completed the first measurement of this web-based study and were included into analyses[2]; 472 of them were female (61.8%), 288 male (37.7%), and four did not specify (0.5%). Ages ranged from 17 to 79 years (M = 48.24, SD = 14.33). The education level of 66.0% of the participants was *Abitur* (German university entrance qualification). On average, the participants had been using the Internet for 18.28

[1] The studies were approved by the ethics committee of the Department 7 of the University of Münster (ID 2018-13-MT). The data of the studies (including full item lists in codebooks) are available at https://doi.org/10.5281/zenodo.3813293.

[2] A total of 13 participants gave no consent, 205 dropped out during answering, and one person was excluded due to just clicking through the survey. There were no significant differences between included and excluded subjects/drop-outs regarding age, gender, educational level, Internet experience or daily surfing time. Of the 292/551 participants at T2/T3, 288/544 could be matched to T1 data.

years (*Min* = 4, *Max* = 35, *SD* = 4.90) and stated an active use of on average 3.33 hours per day (*Min* = 0.2, *Max* = 13, *SD* = 2.18). Participants took part voluntarily and anonymously[3]; they had a chance to win one out of ten €50 vouchers for an online bookshop. At the second time of measurement, n = 292 participants completed the study, n = 551 at the third.

Stimulus material

A pre-study was performed to pre-select a stimulus set that was unknown to participants but still reflected a typical range of general website content quality. To generate this set, we methodically followed Miniukovich and De Angeli (2015). Therefore, N = 64 (22 female, 42 male) participants were recruited in April 2018 via a crowdworking platform. Ages ranged from 19 to 64 years (*M* = 37.64, *SD* = 11.53). The education level of 60.9% of the participants was *Abitur* (German university entrance qualification). On average, the participants had been using the Internet for 16.69 years (*Min* = 5, *Max* = 27, *SD* = 5.27) and stated an active use of on average 6.14 hours per day (*Min* = 1, *Max* = 20, *SD* = 3.62). Participants were asked to find websites that corresponded as closely as possible to ten given categories (Download & Software; E-Commerce; E-Learning; E-Recruiting & E-Assessment; Entertainment; Information site; Presentation & Self-portrayal / Corporate websites; Search engines; Web portals; Weblogs & Social Sharing. Information on the categorization scheme can be found in Thielsch, 2008; p. 86f. and in the Appendix). These websites were required to be written in the German language and could not be among the most visited websites in Germany. Participants had to type in two URLs per category, thus they had to search for 20 websites in total. Participants in the pre-study took part voluntarily and anonymously;

[3] As a supplementary note here and for the following studies: In the survey phase, there was a pseudonymous link for a short time for the purpose of drawing or remuneration. Remuneration was made either by the panel providers themselves or by the researchers via e-mail using general participation codes/hash codes. Since there was no direct assignment or systematic link to the study data and the e-mail addresses were only recorded for the purpose of voucher dispatch/remuneration and were not processed further, all final study data are anonymous.

they were paid €2.25. At the end of the pre-study, participants were thanked and had the opportunity to exclude their data from the subsequent analysis.

This crowdworking approach led to a total of 1280 proposed websites. First, URLs were checked for duplicates. Further, we only considered websites that were written in German and were not part of the 100 most popular websites in Germany, leading to a final pool of 764 websites (available at https://doi.org/10.5281/zenodo.3813293). Out of this pool, we randomly selected one website per category for Study 1 (see Appendix; screenshots can be requested via the corresponding author). A total of 87.2% of the participants in Study 1 stated they were unfamiliar with the presented website.

Measures

To construct a shortened version of the full Web-CLIC, we followed these guidelines: First, each content area contained in the Web-CLIC (Clarity, Likeability, Informativeness, Credibility) should be represented by at least one item. Second, the items should be as representative as possible for their associated content facet. Finally, the resulting scale should be as short as possible so it can be applied when short assessment times are needed. To ensure that each selected item appropriately represented its associated facet, we only considered items that showed high factor loadings in the confirmatory study reported in Thielsch & Hirschfeld (2019, Study 2) and a high item-total correlation in this prior dataset. Next, we only retained those items that were able to most fully describe the associated facet and were easiest to understand. This procedure led to clear item selections for the content facets Likeability, Informativeness, and Credibility; for Clarity two items were equally suitable, so we chose the shorter one. The four selected items are as follows:

(1) Clarity: “The contents of the website are clearly presented.”
(2) Likeability: “I enjoy reading the website.”
(3) Informativeness: “The website is informative.”
(4) Credibility: “I can trust the information on the website.”

Procedure

Three data collection points were planned to measure the short-term stability (after two days) and the medium-term stability (after two weeks) of the Web-CLIC-S. The participants received an invitation for the first data point (T1) via e-mail, sent on April 23, 2018. Two days after T1, the invitation for the second measurement point (T2) was sent, and two weeks after T1 the invitation for the third one (T3) was sent. Every participant evaluated only one website (randomly assigned at T1) at each time of measurement.

Participants were recruited via the German online panel PsyWeb (https://psyweb.uni-muenster.de/). Participation in this panel is completely voluntarily, and members agree on receiving invitations for scientific studies; they can unsubscribe and delete their personal data at any time. At T1, participants of the present study received an invitation to take part in a study about the evaluation of websites over time. Following the invitation link, they were informed about the involved researchers, anonymity, voluntariness, procedure and duration of the study. After being asked for demographic information (age, gender, education level, Internet experience), participants were randomly assigned to one website from the stimulus set. The fully functional website they were assigned was opened in a new window. Participants were instructed to explore the given website and to open some subpages (i.e., the task was free exploration). When returning to the survey window, they were asked whether they were familiar with the website before the study. Then, the shortened version and the remaining items of the Web-CLIC were presented, as well as four validation measures (PWU, VisAWI-S, intention to revisit, overall website score; see Study 2 for a detailed description). Web-CLIC, PWU and VisAWI-S were given in a random order, and all items within the questionnaires were also randomized to avoid any position effects. Afterwards, intention to revisit was rated, and an overall website score was assigned. At the end, participants could comment on the study, they were thanked, and had the opportunity to exclude their data from the subsequent analysis. Participants needed about eight to nine minutes to complete T1.

At both T2 and T3, participants were given a short introduction including a reminder about the study. The participants were again asked for consent, whether they remembered the website presented at T1, and whether they had visited it in the meantime. Afterwards, they were asked to again open the website assigned at T1 (using the same instruction as in T1). When returning to the survey window, participants answered the Web-CLIC-S and gave an overall website score. At the end of each data collection point, participants had the opportunity to exclude their data from subsequent analysis and to give additional comments. On average, participants needed about four to five minutes to complete each measurement. Additionally, at the end of T3, they received additional information about the research and were given the opportunity participate in the lottery.

2.2. Results and discussion

We used the data from T1 and T3 to analyze dimensionality and reliability. As T2 data did not fulfill sample size requirements for calculating a CFA, data from T2 were only used to estimate the short-term retest reliability and internal consistency. First correlations with divergent and concurrent validation measures are reported in Table 2 but were analyzed in Detail in Study 2.

Dimensionality

The proposed unidimensional structure was tested using Velicer's MAP-tests and CFA for T1 and T3. The MAP-test indicated that the optimal number of factors was one. To estimate the model parameters, we used DWLS estimation. Model fit was deemed acceptable if *CFI* and *TLI* >.95 and *RMSEA* <.08 (Hu & Bentler, 1999). The model fit the proposed structure very well at T1 (*CFI* = .99; *TLI* = .98; *RMSEA* = .075) and T3 (*CFI* = .99; *TLI* = .99; *RMSEA* = .036). All items showed large (at least .69) and statistically significant loadings on the proposed g-factor (see Table 1).

Table 1. Item-level results from Study 1.

Item	Mean	SD	Item-scale correlation	Alpha if item deleted	Standardized loadings
The contents of the website are clearly presented	4.89 / 4.91	1.50 / 1.48	.68 / .77	.83 / .85	.69 / .79
I enjoy reading the site	3.65 / 3.73	1.70 / 1.64	.77 / .79	.79 / .85	.79 / .80
The website is informative	4.66 / 4.70	1.60 / 1.51	.80 / .85	.79 / .82	.81 / .86
I can trust information on the website	4.25 / 4.35	1.45 / 1.43	.75 / .76	.81 / .86	.76 / .76

Note: values before / after the slash refer to results from T1 (N = 764) / T3 (N = 551)

Internal consistency and retest reliability

Internal consistency is often considered an indicator of reliability. Thus, we calculated Cronbach's α for the Web-CLIC-S based on the data gathered at T1, T2, and T3. Given the guidelines of Nunnally (1978), Cronbach's α values above .8 can be considered as good, above .9 as excellent. For the Web-CLIC-S, Cronbach's α was .85 in T1, .89 in T2, and .88 in T3. Thus, the Web-CLIC-S exhibited good internal consistencies. Given the shortness of the scale, this outcome is notable.

While the internal consistency can give an impression about homogeneity of a scale and accuracy of item configuration, retest reliability can be interpreted in terms of stability of a measure. To calculate short-term stability, participants again completed the Web-CLIC-S two days after T1, and medium-term stability was assessed by having them take it again two weeks after T1. Retest reliability is interpreted according to the time span between measures, where values above .8 are considered good, values above .7 as sufficient, and values above .6 as acceptable for research purposes and for analyses on group level (Nunnally, 1978). Results for the Web-CLIC-S show good retest values for short-term stability ($r_{T1–T2}$ = .85; 95%CI: .82 - .88; n = 287) and sufficient to good retest

values for medium-term stability ($r_{T1–T3}$ = .79; 95%CI: .76 - .82; n = 542). Thus, the Web-CLIC-S appears to be a stable measure, at least over short and medium periods.

3. STUDY 2: CONSTRUCT VALIDITY OF THE WEB-CLIC-S

The purpose of Study 2 was to evaluate the Web-CLIC-S using several validation strategies, such as examining convergent validity (high correlations with related constructs), divergent validity (low to no connections with unrelated criteria), discriminative validity (for the Web-CLIC-S, the ability to distinguish between different websites), and concurrent validity (correlations with a simultaneously assessed criterion).

3.1. Method

Participants

A total of 341 participants completed this web-based study and were included into analyses[4]; 212 were female (62.2%), 127 male (37.2%), two did not specify (0.6%). Ages ranged from 18 to 93 years (*M* = 42.30, *SD* = 15.87). The education level of 73.6% of the participants was *Abitur* (German university entrance qualification) or higher. On average, the participants had been using the Internet for 17.33 years (*Min* = 3, *Max* = 35, *SD* = 5.37) and stated an active use of on average 3.73 hours per day (*Min* = 0.5, *Max* = 16, *SD* = 2.38). Participants took part voluntarily and anonymously; they had the chance to win one out of five €50 vouchers for an online bookshop.

Stimulus material

Based on the URLs gathered in the pre-study to Study 1, a set of 20 websites from ten different content domains was randomly selected (websites used in Study 1 were

[4]Five participants gave no consent, 197 dropped out during answering, 36 were excluded (mostly based on the JavaScript controlling task fulfilment, see procedure). There were no significant differences between included and excluded subjects/drop-outs regarding age, gender, educational level, Internet experience or daily surfing time.

excluded, see Appendix; screenshots can be requested via the corresponding author). These websites were selected to represent a broad range of corporate and institutional websites in Germany, which make up a large percentage of a person's everyday online activities. Each website category was represented by two websites (see Appendix). A total of 90.0% of the participants stated they were unfamiliar with the presented website.

Measures

For construct validation of the Web-CLIC-S, we used several established measures. Unless otherwise specified, participants were asked to indicate their level of agreement with each item of these questionnaires on seven-point Likert scales ranging from 1 ("strongly disagree") to 7 ("strongly agree"):

Credibility (scale from Appelman and Sundar, 2016): This scale consists of three items and showed good reliability (α = .87) as well as content, criterion, and construct validity (see Appelman & Sundar, 2016, p. 72). We used it as a criterion for convergent validity.

Informativeness and entertainment (Kang & Kim, 2006): Two single items from the main study by Kang and Kim (2006) were used (informativeness: "This website is a valuable resource."; entertainment: "This web site is fun to explore."). Kang and Kim (2006) provided evidence for reliability and discriminant validity of their measure. In the current study, it is used as a criterion for convergent validity of the Web-CLIC-S.

Understandability (AIM-Q, Lee et al., 2002). The AIMQ questionnaire from Lee et al. (2002) is a widely used measure for information quality. We used its understandability scale as a criterion for convergent validity of the Web-CLIC-S. Lee et al. (2002) report a Cronbach's α of .90 for this four-item scale.

Mood: Mood was measured with a graphical five-point smiley scale ranging from a very sad to a very happy smiley face (Jäger, 2004). In a series of two studies, Jäger (2004) provided evidence for unidimensionality and equidistance of this scale as well as

high correlations with the German version of the PANAS scale ($.75 \leq r \leq .89$). Mood is used as a criterion for divergent validity.

Perceived website aesthetics (VisAWI-S, Moshagen & Thielsch, 2013): The short version of the Visual Aesthetics of Websites Inventory (Moshagen and Thielsch, 2013) was used to measure the general factor *subjective aesthetics.* The authors report Cronbach's α values between .76 and .81 for this four-item scale, and a correlation of .91 to the full VisAWI. Evidence for convergent, divergent, and concurrent validity is provided (Moshagen and Thielsch, 2013). The VisAWI-S is used as a criterion for divergent validity.

Perceived website usability (PWU): This one-dimensional scale, measuring perceived website usability, was adapted to German based on Flavián et al. (2006). The PWU is a seven-item measure assessing perceived ease of use, ease of understanding and speed of information retrieval (see Thielsch, 2008; Thielsch et al., 2015). Thielsch (2008) found a Cronbach's α of .95 for the adapted version and provided evidence for factor and convergent validity. The PWU is used as a criterion for divergent validity.

Perceived website usability (UMUX-Lite, Lewis et al., 2013): The UMUX-Lite is a two-item measure based on the System Usability Scale (Brooke, 1996). Lewis et al. (2013) report Cronbach's α values between .82 and .83, concurrent validities of $r = .81$ with the full SUS as well as high correlations for willingness to recommend ($.73 \leq r \leq .74$). The UMUX-Lite is used as a criterion for divergent validity.

Intention to revisit: The four items created for Study 5 from Moshagen and Thielsch (2010) were used to assess participants' intention to revisit the website ("I will visit the website again", "I will visit the website on a regular basis", "I would recommend the website to my friends", "If I have interest in such topics in the future, I would consider visiting this website again"). The responses to these items were averaged to form an index of the participants' intentions to revisit the website. This index is used as a criterion for concurrent validity.

Overall website score: The overall website impression was assessed with a grade on a on a six-point grading scale ("Altogether: I would mark the website with…", 1 = "very good", 2 = "good", 3 = "satisfactory", 4 = "adequate", 5 = "poor", 6 = "unsatisfactory") commonly used in the German education system. This grade was used as a criterion for concurrent validity.

Procedure

Participants were invited via e-mail through the online panel PsyWeb (participants of Study 1 were excluded automatically) and via a student newsletter. The study was announced as general website evaluation study. On the first two survey pages, all participants were informed about involved researchers, anonymity, voluntariness and duration of the study. After being asked for demographic information (age, gender, education level, Internet experience), participants were asked about their mood and then randomly assigned to one website from the stimulus set. The fully functional website in question was opened in a new window. Participants were given the same free exploration task as in Study 1; a JavaScript was used to control if participants were fulfilling the task of opening a website. When returning to the survey window, they were asked if they already knew the website before the study. Then, the shortened version and the remaining items of the Web-CLIC were presented, as well as six validation measures: Credibility scale of Appelman and Sundar (2016), Informativeness and entertainment items of Kang & Kim (2006), the understandability scale from the AIM-Q, VisAWI-S, PWU, and UMUX-Lite. All measures were given in random order, and items within the questionnaires were also randomized to avoid any position effects. Afterwards, intention to revisit and the overall website score were assessed. At the end, participants were thanked and had the opportunity to exclude their data from the subsequent analysis. Finally, participants received additional information about this research and were given the opportunity to participate in the lottery. The study was available online from 16 July 2018 until 30 August 2018; completing it on average took about nine to twelve minutes.

3.2. Results and discussion

Correlations between the Web-CLIC-S and the convergent, divergent and concurrent criteria are shown in Table 2 (which also includes the respective values from Studies 1 and 3). As expected, the Web-CLIC-S sum score showed high correlations with convergent criteria. In particular, very high correlations were found between the Web-CLIC-S and the full Web-CLIC ($.951 \leq r \leq .972$, all $p < .001$). Correlations with credibility, informativeness and entertainment measures were high as well. The comparatively lowest correlation was found with the understandability scale of the AIMQ ($r = .598$, $p < .001$).

Divergent validity refers to the degree to which the instrument is distinct from scales that assess other aspects of subjective perceptions. This was clearly the case when looking at the connection between mood (measured before using the website) and the Web-CLIC-S ($r = .041$, $p = .454$). However, the Web-CLIC-S showed high correlations with measures of aesthetics and usability and these were only slightly below the convergent measures. This pattern was previously observed in the validation of the full Web-CLIC (Study 4 in Thielsch & Hirschfeld, 2019), where it was argued to be in line with the interpretation by Moshagen and Thielsch (2010, p. 701) that good designers strive to jointly optimize content, usability and aesthetics. Particularly, the clarity of website content can support usability (e.g., well-structured and comprehensible content may help users navigate the website), and an aesthetically pleasing website might enhance the likeability of web content. Thus, usability and aesthetics should not necessarily be treated as completely divergent constructs. Still, given such mixed results, additional analysis and proof of Web-CLIC-S's usefulness seem necessary (see Study 3).

The Web-CLIC-S highly correlated with concurrent measures, especially with the overall website score ($.822 \leq r \leq .824$, all $p < .001$). These results even exceed results for the full version of the Web-CLIC and are in line with prior research, stressing that perceptions of website content are important for users' overall attitudes and satisfaction (e.g., De Wulf et al., 2006; McKinney et al., 2002; Shukla et al., 2010), their intention to

revisit, and loyalty (e.g., Aranyi & van Schaik, 2016; Kim & Niehm, 2009; Thielsch et al., 2014).

Table 2: Correlations between the Web-CLIC-S and the convergent, divergent, and concurrent criteria

	Study 1 (N=764)	Study 2 (N=341)	Study 3 (N=309)
Convergent measures			
Web-CLIC full	.972	.951	.949
Credibility		.750	
Informativeness		.768	
Entertainment		.717	
Understandability (AIMQ)		.598	
Divergent measures			
Mood		.041	.085
Disposition to trust			.106
Perceived aesthetics (VisAWI-S)	.736	.765	.721
Perceived usability (PWU)	.663	.723	.668
Perceived usability (UMUX-Lite)		.721	.684
Concurrent measures			
Intention to revisit	.762	.760	.750
Trust in organization			.711
Overall website score	.822	.824	.762

Note: Except for mood ($p = .454$ in Study 2, $p < .01$ in Study 3) and disposition to trust ($p < .01$), all correlations are significant with $p < .001$. Overall website score was recoded so that high Web-CLIC-S values correspond to high overall scores. Correlations for Study 3 are based on a random-effect meta-analysis.

Finally, we analyzed the discriminative validity of the Web-CLIC-S, i.e., the ability of the measure to distinguish between different websites. To test whether the Web-CLIC-S score differ as a function of the given website, an ANOVA was calculated (dependent variables: Web-CLIC-S; independent variable: evaluated website). The ANOVA was

significant, $F = 11.317$, $df = 19/321$, $p < .001$, $\eta^2 = .401$, indicating that different websites received clearly different evaluations on the Web-CLIC-S. When comparing the most negatively evaluated website (a Download & Software website) with the most positively evaluated one (an E-Learning website), a highly significant difference emerged (t (29) = 8.85, $p < .001$, $d = 3.191$), meaning that those two websites differed on the Web-CLIC-S by more than three standard deviations. Thus, we conclude that the Web-CLIC-S is very capable of discriminating between different websites.

4. STUDY 3: USEFULNESS OF THE WEB-CLIC-S

The main goals of Study 3 were to perform additional validations and to demonstrate the usefulness of the Web-CLIC-S. Regarding the first aim, we provide further evidence for the validity of the Web-CLIC-S by comparing it to validation scales as done in Study 2 as well as to additional measures. Given the partly mixed results in Study 2, we focused on divergent and concurrent measures. Regarding the second aim, we tested whether Web-CLIC-S ratings are related to actual behavior: donations to one of three different organizations (i.e., predictive validity). In addition, we tested whether the Web-CLIC-S explained variance above and beyond other website ratings.

4.1. Method

Participants

A total of 309 participants completed this web-based study and were included into analyses[5]; 149 were female (48.2%), 157 male (50.8%), three did not specify (1.0%). Ages ranged from 18 to 82 years ($M = 48.25$, $SD = 15.66$). The education level of 69.9%

[5]Two participants gave no consent, 143 dropped out during answering, 42 were excluded (mostly based on the JavaScript controlling task fulfilment, see procedure). There were no significant differences between included and excluded subjects/drop-outs regarding age, gender, educational level, Internet experience or daily surfing time.

of the participants was *Abitur* (German university entrance qualification) or higher. On average, the participants had been using the Internet for 18.16 years (*Min* = 5, *Max* = 32, *SD* = 5.11) and stated an active use of on average 2.43 hours per day (*Min* = 0.15, *Max* = 12, *SD* = 1.76). The participants took part voluntarily and anonymously; as compensation, they received €8 in the form of a voucher for an online bookstore.

Stimulus materials and measures

We used a within-subject design with three stimuli. All subjects rated the same three websites of non-profit organizations in a random order. Using search engines, we selected typical organizations that support people with depression and provide general information about the disease. Tested websites were from the initiatives "Deutsche DepressionsLiga e.V." (https://www.depressionsliga.de/), "Stiftung Deutsche Depressionshilfe" (https://www.deutsche-depressionshilfe.de) and "Bundesverband für Gesundheitsinformation und Verbraucherschutz - Info Gesundheit e.V. (BGV)" (https://bgv-depression.de/; screenshots can be requested via the corresponding author).

As a measuring instrument, the shortened version (and the remaining items of the full version) of the Web-CLIC were presented, as were six additional measures: Mood, PWU, UMUX-Lite, VisAWI-S, intention to revisit, and the overall website score (for descriptions of measures, see Study 2). Furthermore, participants rated their trust in each organization using the trusting belief scales of McKnight, Choudhury and Kacmar (2002). The authors divide trust beliefs into three factors: competence (four items), benevolence (three items), and integrity (four items); with this classification, they follow the very established trust model of Mayer, Davis and Schoorman (1995). McKnight and colleagues (2002) report very good reliability values in the sense of internal consistency (per scale $.91 \leq \alpha \leq .95$). They also applied different validation strategies and found evidence for convergent and discriminant validity (McKnight et al., 2002). However, the assumed factorial structure of the three scales could not be confirmed perfectly in their original study. Thus, in the present analysis, we used the overall score of this measure. In addition, we assessed participants' general disposition to trust websites using an adoption of the trusting stance scale of McKnight et al. (2011). This scale consists of three items

and has shown good reliability (α = .86). Except for the overall grade, participants were asked to indicate their level of agreement with each item on seven-point Likert scales ranging from 1 (“strongly disagree”) to 7 (“strongly agree”).

Procedure

Participants were invited via e-mail through the online panel PsyWeb; participants of prior Web-CLIC-S studies were excluded automatically. The survey was announced as a study on the evaluation of health websites. On the first two survey pages, all participants were informed about the involved researchers, anonymity, voluntariness, the procedure, and the duration of the study. After they gave demographic data, as in previous studies, participants were asked about their general disposition to trust websites, their donation behavior over the last 12 months, their familiarity with depression and the three organizations whose websites were tested. Then, the three fully functional websites were randomly presented in a new window along with a free exploration task, as in Studies 1 and 2. A JavaScript monitored whether participants opened the websites. For each website, participants answered the given measures; after that, participants received a forced-choice item about which of the three organizations should receive a donation of €100 (money was provided by the investigators). In addition, they were asked whether they would donate their €8 study compensation to one of the organizations (all participants received the compensation regardless of the response). Finally, participants had the opportunity to exclude their data from subsequent analyses and to comment on the study. They were thanked and received further information about the background of the research. The study was available online from 10 September 2018 until 12 October 2018; completing it took on average 21 to 24 minutes.

4.2. Results and discussion

First, we analyzed correlations between the Web-CLIC-S and convergent, divergent and concurrent measures (see Table 2). To combine the three different correlations from the different websites into one estimate (rho), we performed a random-effect meta-analysis on the individual correlation coefficients. Again, we found very high correlations

between the Web-CLIC-S and the full Web-CLIC (*rho* = .949, $p < .001$) as well as very high correlations between Web-CLIC-S and concurrent measures, including the newly investigated construct trust in organization (*rho* = .711; $p < .001$). With respect to divergent measures, we found the same pattern as in Study 2 regarding the correlations with usability and aesthetics scales ($.668 \leq rho \leq .721$; $p < .001$). Yet, as in Study 2, correlations were much lower in relation to mood (*rho* = .085; $p < .01$) and the newly assessed disposition to trust (*rho* = .105; $p < .01$). Thus, these results might reflect the typical procedure of web designers to jointly optimize content, usability and aesthetics (as discussed in Study 2), but they also indicate that the Web-CLIC-S ratings were only very weakly influenced by emotional states (such as mood) or users' personality traits (such as disposition to trust).

Second, we wanted to determine whether the website that received the highest global Web-CLIC-S rating was also the website that participants indicated should get the donated money. For this, we determined the website that received the highest Web-CLIC-S rating for each participant (see Table 3). For 59 participants, the highest Web-CLIC-S score was tied, i.e., two websites got a similarly high rating and were thus excluded. For the remaining 250 participants, we used a chi-square test to analyze the association between Web-CLIC-S rating and the forced choice between one of the organizations. The results indicated that there was a significant association between content ratings and the decision about which organization should receive the donated money ($\chi^2 (4) = 77.74$; $p < .001$), showing a "medium effect" (*Cramers V* = .394) according to Cohen's guidelines (Cohen, 1988). Similar results were found for the individual donation decision regarding whether an individual participant would donate their study compensation ($\chi^2 (6) = 42.77$; $p < .001$; *Cramers V* = .292).

Table 3. Relationship between highest content ratings and the decision to donate money for a specific organization (n = 250).

		Donation recipient		
		1)	2)	3)
Highest content rating	1) Deutsche DepressionsLiga e.V.	33	13	3
	2) Stiftung Deutsche Depressionshilfe	39	83	15
	3) Bundesverband für Gesundheitsinformation und Verbraucherschutz-Info Gesundheit e.V.	17	14	33

Note: This table refers to the forced-choice donation decision, since about half of the participants in the second donation question (donation of their own study compensation) decided to keep the money.

Third, we tested whether the Web-CLIC-S explains variance above and beyond name recognition and a simple global item. For this, we used a multiple logistic regression model to predict to which organization money should be donated (using the forced-choice donation item). Our critical comparison involved two models. The first used information about whether a participant recognized the organization and the overall grade they gave the organization's website to predict whether a participant intended to donate money to this organization. The second model used information about the participant's knowledge of the organization, the overall grade they gave it, and the Web-CLIC-S to predict whether a participant intended to donate money to this organization. The two models were compared using likelihood ratio tests, and variance explained was measured using Tjur's D (Tjur, 2009). We found that the second model (*Tjur's D* = .13) showed a small albeit significantly better fit to the data than the first model (*Tjur's D* = .12; χ^2 (1) = 6.32, p = .012). Similarly, the Web-CLIC-S showed a significant amount of incremental validity when comparing it to the UMUX-Lite (χ^2 (1) = 23.17, p < .001), PWU (χ^2 (1) = 21.53, p < .001), or the VisAWI (χ^2 (1) = 22.172, p < .001).

In sum, we demonstrated with this study that the evaluations gathered with the Web-CLIC-S are highly useful in predicting not only intentions but also actual user behavior.

5. STUDY 4: BENCHMARKS AND OPTIMAL CUT POINTS

Most website evaluation tools yield continuous scores, leading to, for example, a website usability score of 5.5 – within themselves, these scores are difficult to interpret. Benchmarks and cut points (see Thielsch et al., 2019) are valuable alternatives when it is not possible to compare website ratings with prior versions of the same website or similar other websites (A/B testing). Both, as described below, allow for meaningful interpretations of individual scores, which can then support decisions made after an evaluation.

Benchmarks enable one to compare websites in a given domain, as they are usually based on a pool of comparable websites. For example, benchmarks may be presented as the means and standard deviations of previous summed website ratings. Thus, using this benchmark, one can assess whether a specific website is perceived as more or less informative than an average site from a given test pool. Yet, 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.

This problem can be tackled by using optimal cut points. Cut points 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 applied on website evaluations (see Hirschfeld & Thielsch, 2015). This procedure is inspired by methods in medicine and needs an external criterion, such as global ratings of users' overall impressions of websites or a website ranking. In contrast to benchmarks, they require a substantial but comparatively smaller sample size, especially if large differences exist between positive and negative stimuli.

Study 4 aims at providing benchmarks and optimal cut points. For the cut point analyses, we combined data from 15 different website evaluation studies: data from the current paper (T1 of Study 1, Study 2, and Study 3), data from Dames et al. (2019), data from Thielsch and Hirschfeld (2019; Study 2, T1 of Study 3, Study 4, and Study 6), data from Thielsch et al. (2019), and data from six additional, currently unpublished studies

from our research group. In these studies, the Web-CLIC-S was applied together with an overall website evaluation, that we used as the criterion for the cut point analyses. For the benchmark analysis, we included additional data from Thielsch and Thielsch (2018), Thielsch and Wirth (2017), and one additional, currently unpublished study from our research group.

5.1. Method

Participants

A combined sample of 9,754 respondents was used for benchmark analysis, among them 5,266 females (54.0%), 4,469 males (45.8%), and 19 who did not specify (0.2%). Of those, data from 7,955 respondents were used in cut point analyses (52.4% females, 47.5% male, 0.1% not specified). Ages ranged from 14 to 93 years (M = 41.62, SD = 16.19 in benchmark analyses and M = 41.59, SD = 16.45 in cut point analyses). The education level of 58.8% of the participants was *Abitur* (German university entrance qualification) or higher (56.0% in cut point analyses); for 5.8% of respondents, specific data on educational level were not available. On average, participants had been using the Internet for 14.88 years (Min = 1, Max = 45, SD = 5.57; data available for n = 9,221) and stated an active use of on average 2.96 hours per day (Min = 0.01, Max = 16, SD = 2.16; data available for n = 9,081). Participants included in the cut point analyses had been using the Internet for 14.74 years (Min = 1, Max = 45, SD = 5.69; data available for n = 7,944) and stated an active use of on average 2.98 hours per day (Min = 0.01, Max = 16, SD = 2.17; data available for n = 7,846). In all studies, participants took part voluntarily and anonymously, meaning we cannot rule out that some participants took part twice (yet, additional cut point analyses with the largest unique sample of 2,265 participants resulted in very similar results, see below). In most studies, participants received no compensation but could request a summary of the study's results; in some studies they received a financial compensation, could take part in a lottery for vouchers, or could receive course credits for participation (when participants were students).

Stimulus material and procedure

In each study, participants were informed about its objective, the involved researchers, anonymity, voluntariness and duration. After providing demographic information, participants were usually randomly assigned to one or two fully functional websites from the respective stimulus set; only in one study were participants asked to evaluate more than three websites. In sum, 12,568 ratings on 175 websites and on additional eight online annual business reports (see Thielsch & Wirth, 2017) were analyzed (in cut point analyses, there were 9,865 ratings on 168 websites). On average, a website was evaluated by 68.68 participants (*Min* = 13, *Max* = 881); in cut point analyses, this was *M* = 58.72 participants (*Min* = 13 and *Max* = 481). All websites except for two could be sorted into one of 12 different categories. In studies included in the cut point analyses, a six-point grading scale (1 = "very good", 2 = "good", 3 = "satisfactory", 4 = "adequate", 5 = "poor", 6 = "unsatisfactory") was applied. At the end of each study, participants could exclude their data from subsequent analysis and were thanked.

5.2. Results and discussion

Influences of age, gender, and education level

Before calculating benchmarks, we first examined the extent to which the response to the Web-CLIC-S was associated with differential factors such as age, gender, or education. The multiple linear regression showed significant effects for age ($b = 0.007$; $p < .001$), gender ($b = 0.122$; $p < 0.001$) and educational level ($b = 0.144$; $p < 0.001$). However, together these factors explained only about 3.6% of the variance in the Web-CLIC-S ratings, indicating that demographic differences are of little practical importance.

In contrast, clear differences appeared in an ANCOVA with the website category as the independent variable, the Web-CLIC-S mean score as the dependent variable, and age, gender, and education level as covariates: $F_{website\ category} = 200.9$, $df = 11/11107$, $p < .001$, $\eta^2 = .16$. We thus decided to calculate benchmark values and optimal cut points based on the different website categories and an overall benchmark, but not to

standardize values psychometrically for different age groups, genders, or levels of education. Still, in specific situations it might be important to monitor such variables, such as when analyzing special target groups or evaluations of specific web contents with relation to age, gender, or education.

Benchmarks for different website categories

We calculated Web-CLIC-S mean and standard deviations separately for each website category. This benchmark (Table 4) can be used to compare results from a newly tested website with the respective category.

Table 4: Benchmarks for the Web-CLIC-S

Category	*M*	*SD*
Download & Software (m = 7; n = 213)	3.55	0.57
E-Commerce (m = 14; n = 1,039)	4.43	0.55
Entertainment (m = 7; n = 301)	3.16	0.39
E-Learning (m = 7; n = 192)	4.54	0.27
E-Recruiting & E-Assessment (m = 26; n = 1,438)	4.53	0.35
Information (m = 12; n = 545)	4.67	0.63
Information: E-Health (m = 37; n = 4,200)	5.06	0.33
Presentation & Self-portrayal: Websites (m = 40; n = 2,197)	4.46	0.69
Presentation & Self-portrayal: Online business reports (m = 8; n = 165)	4.38	0.38
Search engines (m = 6; n = 249)	4.56	0.55
Web portals (m = 8; n = 339)	3.73	0.67
Weblogs & Social Sharing (m = 9; n = 241)	3.62	0.68
Overall score (m = 183; n = 12,568)	4.46	0.71

Note: *M* = mean, *SD* = standard deviation, m = number of evaluated websites in one category, n = number of respondents. Evaluations of online annual business reports were included as a subcategory of the Presentation & Self-portrayal category, representing a special form of typical web-based corporate communications (see Thielsch & Wirth, 2017). The overall score includes the ratings of two websites that could not be placed in one of the categories.

Yet, because participants in most studies were randomly assigned to websites that were unfamiliar to them, the benchmark largely reflects an evaluation by a random web user, not by people highly familiar with a given website (such as registered customers of an e-commerce website). In addition, in some categories fewer than ten websites were tested, meaning that results might be different if a larger set of stimuli were examined. In such cases, or if no category in the benchmark fits at all, we recommend using the general cut points presented in the following section.

Cut point analyses

In order to establish meaningful cut points to interpret the Web-CLIC-S, we used receiver operating characteristic (ROC)-based methods (see Hirschfeld & Thielsch, 2015; Thiele & Hirschfeld, in press). These methods identify the cut points for the content ratings that best differentiate between websites that were overall rated as good (grades 1 or 2) and those that were overall rated as not good (grades 3, 4, 5, or 6). Specifically, these methods entail calculating the sensitivity and specificity of different cut points. The *sensitivity* refers to the percentage of good websites that actually got a scale score larger than the cut point, whereas *specificity* refers to the percentage of bad websites that actually got a scale score smaller than the cut point. The cut point that yields the highest sum of sensitivity and specificity (i.e., Youden index) is identified as optimal. To determine the variability of the cut point and determine whether different cut points should be used for the different website categories, we used bootstrapping with 5,000 repetitions. We found that websites that were overall rated as good received a higher Web-CLIC-S rating (M = 5.42) than websites rated as not good (M = 3.80; t (9863) = -81,964, $p < .001$, d = -1.65).

Furthermore, the Web-CLIC-S showed an area under curve (AUC) of .884 (95% CI: .88 - .89), indicating a good classification of the websites based on the overall rating. The cut point that was defined as optimal was 4.66; namely, Web-CLIC-S ratings below 4.66 indicate a “bad” website, while ratings higher than 4.66 indicate a “good” website. Using this cut point to determine whether a website is generally perceived as good or bad would result in 83 percent of the good websites actually identified as good (sensitivity) and 78

percent of the bad websites actually identified as bad (specificity). Testing the variability of the optimal cut point using bootstrapping resulted in a 95% CI between 4.64 and 4.69. This indicates that we were able to estimate the optimal cut point with a relatively high precision.

When separately estimating optimal cut points for the different website categories, we found that there were systematic differences in the cut points identified as optimal. Entertainment websites had the lowest value for an optimal cut point (*OC* = 3.71; 95% CI = 3.45 – 4.01), and E-health websites had the highest value for an optimal cut point (*OC* = 4.90; 95% CI = 4.85 – 4.98). Table 5 gives an overview of the cut points identified as optimal for the different groups as well as the associated performance measures (AUC, sensitivity, specificity). Upon replicating the cut point analysis in the largest sample with unique participants, we identified an optimal cut point of 4.85 with a 95% CI between 4.77 and 4.91. Since health websites were tested in that particular study, this replicates our results.

Overall, the results indicate that it is feasible to interpret the Web-CLIC-S in a binary way based on the presented cut points. Based on the fairly large sample size, the high agreement between bootstrapping samples indicates a high stability of these cut points.

Table 5: Optimal cut points for the Web-CLIC-S, including AUC, sensitivity, and specificity.

Category	Optimal cut point	AUC	Sensitivity	Specificity
Download & Software (m = 7; n = 213)	4.208 [4.026-4.405]	0.752	0.661	0.821
E-Commerce (m = 14; n = 1,039)	4.660 [4.547-4.783]	0.859	0.811	0.761
Entertainment (m = 7; n = 301)	3.707 [3.447-3.982]	0.788	0.673	0.797
E-Learning (m = 7; n = 192)	4.723 [4.54-4.925]	0.765	0.747	0.699
E-Recruiting & E-Assessment (m =26; n = 1,434)	4.635 [4.541-4.725]	0.873	0.81	0.792
Information (m = 11; n = 468)	4.512 [4.310-4.720]	0.869	0.774	0.813
Information: E-Health (m = 33; n=3,192)	4.904 [4.853-4.981]	0.875	0.819	0.758
Presentation & Self-portrayal: Websites (m = 40; n=2,197)	4.577 [4.464-4.691]	0.892	0.815	0.798
Search engines (m = 6; n=249)	4.638 [4.385-4.877]	0.854	0.746	0.786
Web portals (m = 8; n=339)	4.458 [4.254-4.679]	0.785	0.722	0.809
Weblogs & Social Sharing (m = 9; n=241)	4.359 [4.258-4.484]	0.914	0.935	0.821

Note: *M* = mean, *SD* = standard deviation, m = number of evaluated websites in one category, n = number of participants. For cut point analysis, data from 11 website categories were available.

6. GENERAL DISCUSSION

Given the importance of content perceptions and the need to measure them concisely and quickly, we successfully developed and validated a short version of the Web-CLIC. The Web-CLIC-S reflects the g-factor of subjective web content experience. In Study 1, we demonstrated the unidimensionality and high reliability of the Web-CLIC-S, including the stability of results over medium periods of time. In terms of reliability, the results for the short version were only slightly lower than for the full version (Thielsch & Hirschfeld, 2019), which is remarkable given the brevity of the scale. The same was true for the various indicators of validity, which are also comparable with the full version and even partly exceed the full version in terms of concurrent validity.[6]

In three studies, high correlations were found between the Web-CLIC-S and the full Web-CLIC ($.949 \leq r \leq .972$, all $p < .001$). As expected, the Web-CLIC-S showed high correlations with convergent and concurrent measures. While Study 2 demonstrated that the Web-CLIC-S is capable of discriminating between different websites, it also showed high correlations with the theoretically rather divergent constructs usability and aesthetics. Thus, following the assumption by Moshagen and Thielsch (2010, p. 701) that good designers usually optimize content, usability and aesthetics at once, we tested additional divergent measures. In doing so, we found that the Web-CLIC-S is not or is only very weakly influenced by emotional states (such as mood, see Studies 2 and 3), users' personality traits (such as disposition to trust, see Study 3), or basic user demographics (see Study 4); thus, the Web-CLIC-S also showed divergent validity and robustness to bias factors. In these studies, instead of simply relying on the validity evidence for the full version, which is a typical mistake in short-form development (Smith et al., 2000), we carefully gathered novel evidence for the Web-CLIC-S that in some regards exceeded the full version's performance on corresponding tests. Finally, practical utility of the Web-CLIC-S was demonstrated by its capability in predicting user

[6] The evaluation of psychometric criteria focused on reliability and validity. Since objectivity is a necessary condition for reliability, positive evaluations in terms of reliability also indicate high objectivity. Moreover, with the Web-CLIC-S being a standardized measure, objectivity in the test situation can easily be achieved, especially when it is carried out in a computer-based manner.

intentions and actual user behavior with incremental validity above and beyond other website measures (see Study 3).

6.1. Interpretation of the Web-CLIC-S and practical implications

The Web-CLIC-S is a very short measure. After exploring a website, most people need less than a minute to answer the four items. Additionally, the items are easy to understand, no specific knowledge or expertise and almost no instruction is needed (as instruction in our studies, we just asked participants to rate a given website). In our studies, we presented the Web-CLIC-S items with a Likert scale ranging from 1 (totally disagree) to 7 (totally agree), where all anchor points were verbally labeled (see online supplements). We found no evidence for systematic differences between different age-groups, genders or educational-levels (see Study 4). The Web-CLIC-S could be used to survey adults and adolescents (older than 14 years), as we have no experience with its application in studies with children. In practical use, we recommend testing a fully functional version of the website in question and having users perform relevant tasks (e.g., searching or browsing tasks, see Dames et al., 2019) to simulate typical use. In general, the wording of items should not be changed, except for minor adjustments to ensure comprehensibility and a perfect fit to the target stimulus. However, no item should be completely removed, as the Web-CLIC-S is already very short, and any further reduction will probably compromise its psychometric quality.

When the user survey is done, the overall mean of the Web-CLIC-S can be calculated by adding up the single values of each item and dividing the resulting sum by four. The Web-CLIC-S should only be interpreted on this scale mean level and not on the single-item level. If specific facets of website content are of interest, the full Web-CLIC or its respective scale (Thielsch & Hirschfeld, 2019) should be applied.

When interpreting Web-CLIC-S mean values, it is essential to consider the subjective character of such an evaluation, as it is not an algorithmic measure of website content. A high value on the scale does not indicate that the content of the website is particularly well made, but rather that website users had a *positive perception* of the content. For

practical interpretation, we determined an optimal cut point for the Web-CLIC-S, indicating that values above 4.66 are desirable. Specific cut points for different content domains as can be found in Table 5. Additional benchmark values for 12 different content domains of websites can be found in Table 4.

6.2. Limitations and future research

Some limitations must be considered when interpreting the present findings, and some even highlight possible avenues for future research. First, as with the full version, the Web-CLIC-S is limited to evaluating subjective content perceptions. For future research, it might be interesting to combine such subjective measures of content with results from automatic algorithms.

Second, the Web-CLIC-S was built on the widely tested Web-CLIC, and its construction included more than 1,400 participants who evaluated 33 websites from a variety of domains. But, neither the tested websites nor the participants are perfectly representative of the enormous number of existing websites and web users. Thus, we cordially welcome replications of our studies and further investigations of the validity and applicability of our measure.

Third, all tested participants shared a common cultural background, as studies were conducted in German. While the full Web-CLIC was successfully applied in an English version in the study by Dames and colleagues (2019), such a test is lacking for the Web-CLIC-S. Culture is a possible cause of bias, as it plays an important role in website content (see Fletcher, 2006), and cultural differences are even found on the level of content features (e.g., Robbins & Stylianou, 2003; Zhao et al., 2003). Thus, future studies should investigate cultural effects of subjective content perceptions as well as possible effects on the Web-CLIC-S. As for the full version, different language versions of the short measure are highly welcome. When performing such research, it would be important to test whether cultural differences are due to how the measure operates in different cultural contexts (i.e., lack of measurement invariance) or whether they are due to real differences in how website content is perceived in different cultures.

6.3. Conclusion

The present research focused on validating a screening tool for subjective perceptions of web content. In extensive psychometric tests, the shortened version of the Web-CLIC showed high reliability and construct validity. Furthermore, as for the full version, the Web-CLIC-S could predict user intentions and behavior. Consequently, we can recommend the use of the measure in future research, and we have provided additional interpretation aids, such as optimal cut points and benchmarks, to facilitate its application in practice. In sum, when a short scale for web content evaluation is needed, the Web-CLIC-S is a sound measure of high value, allowing for a valid screening of users' subjective content perceptions in all situations that require short survey times.

Acknowledgments. We thank Simon Eisbach for his support in conducting Studies 1 and 2 and for providing the JavaScript for task control. We also thank David Kahre for his help with Study 3 and the benchmarking studies. Finally, the authors are grateful for the award of a grant supporting the sampling of Study 3 and an additional study that was included in the benchmark and cut point analysis: This research is supported by the Federal Centre for Health Education (BZgA) on behalf of the Federal Ministry of Health.

Declaration of interest statement

The authors declare that they have no competing interests.

REFERENCES

Ahn, T., Ryu, S., & Han, I. (2007). The impact of Web quality and playfulness on user acceptance of online retailing. *Information and Management, 44*(3), 263–275. doi:10.1016/j.im.2006.12.008

Appelman, A., & Sundar, S. S. (2016). Measuring Message Credibility: Construction and Validation of an Exclusive Scale. *Journalism & Mass Communication Quarterly*, *93*(1), 59–79. doi:10.1177/1077699015606057

Aranyi, G., & van Schaik, P. (2016). Testing a model of user-experience with news websites. *Journal of the Association for Information Science and Technology*, *67*(7), 1555–1575. doi:10.1002/asi.23462

Bansal, G., Zahedi, F. M., & Gefen, D. (2010). The impact of personal dispositions on information sensitivity, privacy concern and trust in disclosing health information online. *Decision Support Systems*, *49*(2), 138–150. doi:10.1016/j.dss.2010.01.010

Brooke, J. (1996). SUS - A quick and dirty usability scale. *Usability Evaluation in Industry*, *189*(194), 4–7. doi:10.1002/hbm.20701

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. Hillsdale: Erlbaum.

Dames, H., Hirschfeld, G., Sackmann, T. & Thielsch, M. T. (2019). Searching vs. Browsing - The influence of consumers' goal directedness on website evaluations. *Interacting with Computers, 31*(1), 95-112. doi:10.1093/iwc/iwz006

De Wulf, K., Schillewaert, N., Muylle, S., & Rangarajan, D. (2006). The role of pleasure in web site success. *Information & Management, 43*(4), 434–446. doi:10.1016/j.im.2005.10.005

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

Fletcher, R. (2006). The impact of culture on web site content, design, and structure: An international and a multicultural perspective. *Journal of Communication Management*, 10(3), 259–273. doi:10.1108/13632540610681158

Hirschfeld, G. & Thielsch, M. T. (2015). Establishing meaningful cut points for online user ratings. *Ergonomics, 58* (2), 310-320. doi:10.1080/00140139.2014.965228

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. doi:10.1080/10705519909540118

ISO (2006). *ISO 9241: Ergonomics of Human-System Interaction – Part 151: Guidance on World Wide Web Interfaces*. Geneva: International Organization for Standardisation.

Jäger, R. (2004), „Konstruktion einer Ratingskala mit Smilies als symbolische marken [Construction of a rating scale with smilies as symbolic labels]", *Diagnostica,* 50, (1), 31-38.

Kang, Y., & Kim, Y. (2006). Do visitors' interest level and perceived quantity of web page content matter in shaping the attitude toward a web site? *Decision Support Systems, 42*(2), 1187–1202. doi:10.1016/j.dss.2005.10.004

Kim, H., & Niehm, L. S. (2009). The Impact of Website Quality on Information Quality, Value, and Loyalty Intentions in Apparel Retailing. *Journal of Interactive Marketing, 23*(3), 221–233. doi:10.1016/j.intmar.2009.04.009

Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: A methodology for information quality assessment. *Information and Management*, *40*(2), 133–146. doi:10.1016/S0378-7206(02)00043-5

Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013, April). 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.

Liu, C., White, R. W., & Dumais, S. (2010). Understanding web browsing behaviors through Weibull analysis of dwell time. *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '10* (379-386. doi:10.1145/1835449.1835513

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Journal*, *20*(3), 709–734.

McKinney, V., Yoon, K., & Zahedi, F. (2002). The measurement of Web-customer satisfaction: An expectation and disconfirmation approach. *Information Systems Research, 13*(3), 296–315. doi:10.1287/isre.13.3.296.76

McKnight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology. *ACM Transactions on Management Information Systems*, *2*(2), 1–25. doi:10.1145/1985347.1985353

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, *13*(3), 334–359. doi:10.1287/isre.13.3.334.81

Meeßen, S., Thielsch, M. T., Riehle, D. & Hertel, G. (2020). Trust is Essential: Positive Effects of Information Systems on Users' Memory require Trust in the System. *Ergonomics, 63* (7), 909-926. doi:10.1080/00140139.2020.1758797

Moshagen, M. & Thielsch, M. T. (2010). Facets of visual aesthetics. *International Journal of Human-Computer Studies, 68* (10), 689-709. doi:10.1016/j.ijhcs.2010.05.006

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

Miniukovich, A., & De Angeli, A. (2015, April). Computation of interface aesthetics. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1163-1172).

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Robbins, S. S., & Stylianou, A. C. (2003). Global corporate web sites: an empirical investigation of content and design. *Information & Management, 40*(3), 205–212. doi:10.1016/S0378-7206(02)00002-2

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

Shukla, A., Sharma, N. K., & Swami, S. (2010). Website characteristics, user characteristics and purchase intention: mediating role of website satisfaction. *International Journal of Internet Marketing and Advertising, 6*(2), 142. doi:10.1504/IJIMA.2010.032479

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment*, *12*(1), 102–111. doi:10.1037/1040-3590.12.1.102

Thiele, C. & Hirschfeld, G. (in press). Cutpointr: Improved estimation and validation of optimal cutpoints in R. Journal of Statistical Software.

Thielsch, M.T. (2008). *Ästhetik von Websites [Aesthetics of websites]*. Münster: MV Wissenschaft.

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

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

Thielsch, M. T. & Thielsch, C. (2018). Depressive symptoms and web user experience. *PeerJ* 6:e4439. doi:10.7717/peerj.4439

Thielsch, M. T., Thielsch, C. & Hirschfeld, G. (2019). How informative is informative? Benchmarks and optimal cut points for E-Health Websites. *Mensch und Computer 2019 – Workshopband* (S. 448-452). Bonn: Gesellschaft für Informatik e.V.. doi:10.18420/muc2019-ws-642

Thielsch, M. T. & Wirth, M. (2017). Web-based annual reports at first contact: corporate image and aesthetics. *Technical Communication, 64* (4), 282-296.

Tjur, T. (2009). Coefficients of Determination in Logistic Regression Models—A New Proposal: The Coefficient of Discrimination. *American Statistician*, *63*(4), 366–372. doi:10.1198/tast.2009.08210

Zhao, W., Massey, B., Murphy, J., & Fang, L. (2003). Cultural Dimensions of Website Design and Content. *Prometheus, 21*(1), 74–84. doi:10.1080/0810902032000051027

APPENDIX: STIMULI

URLs of websites tested in Studies 1, 2, and 3.

Website category	Definition of category	Website URLs study 1	Website URLs study 2	Website URLs study 3
Download & Software	Websites providing free or fee-based apps, programs or codes for downloads.	https:// www.shareware.de	http://www.winsoftware.de https://www.nchsoftware.com/de	
E-Commerce	Websites with the primary aim of buying and selling.	https://www.doorout.com	https://www.alternate.de https://www.kalaydo.de	
E-Learning	Online learning content and webpages for learning.	https://www.udemy.com	https://de.babbel.com https://online-lernen.levrai.de	
E-Recruiting & E-Assessment	Web-based recruiting and assessment.	https://de.indeed.com	https://www.stepstone.de https://www.monster.de	
Entertainment	Websites with the main aim to entertain	https://www.langweiledich.net	https://www.twitterperlen.de https://poki.de	
Information site	Websites with a strong focus on information (also containing passive use of weblogs and wikis).	http://www.rhein-angeln.de	https://www.visitberlin.de/de https://www.einfachlebenretten.de	https://www.depressionsliga.de https://www.deutsche-depressionshilfe.de https://bgv-depression.de
Presentation & Self-portrayal (corporate websites)	Websites of institutions, organizations, and companies for representation and image cultivation	https://www.buenting.de	https://www.hochschule-rhein-waal.de/de https://de.werfen.com	
Search engines	Websites serving for the search of other websites, products, services or the like.	https://duckduckgo.com	https://www.unbubble.eu https://www.ecosia.org	
Web portals	Websites providing an overview of many different issues, offering information and additional links and services.	https://www.freenet.de	http://www.prcenter.de https://www.aol.de	
Weblogs and Social Sharing	Websites serving for creation of virtual chronological diaries, collaborative text editing, immediate networking and interaction of the users or for sharing of resources (e.g., pictures, links, video)	https://nebenan.de	https://www.fanfiktion.de https://www.forum-fuer-senioren.de	

Note. Fully functional websites were linked with the named URL; screenshots can be requested via the corresponding author.

The Web-CLIC-S in German

Bitte beurteilen Sie den Inhalt der Ihnen vorliegenden Website anhand der folgenden Aussagen auf einer Skala von 1 (stimme gar nicht zu) bis 7 (stimme voll zu). Vielen Dank!

	Stimme gar nicht zu	*Stimme nicht zu*	*Stimme eher nicht zu*	*neutral*	*Stimme eher zu*	*Stimme zu*	*Stimme voll zu*
Die Inhalte sind anschaulich aufbereitet.	①	②	③	④	⑤	⑥	⑦
Ich lese diese Website gerne.	①	②	③	④	⑤	⑥	⑦
Die Website ist informativ.	①	②	③	④	⑤	⑥	⑦
Ich kann den Informationen auf der Website vertrauen.	①	②	③	④	⑤	⑥	⑦

The Web-CLIC-S in English

Please judge the content of present website according to the following statements on a scale ranging from 1 (strongly disagree) to 7 (strongly agree). Thank you very much!

	Strongly disagree	Disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Agree	Strongly agree
The contents of the website are clearly presented.	①	②	③	④	⑤	⑥	⑦
I enjoy reading the website.	①	②	③	④	⑤	⑥	⑦
The website is informative.	①	②	③	④	⑤	⑥	⑦
I can trust the information on the website.	①	②	③	④	⑤	⑥	⑦