Quality and User Experience (2023) 8:5
https://doi.org/10.1007/s41233-023-00058-3

RESEARCH ARTICLE

Do you agree? Contrasting Google's Core Web Vitals and the impact of cookie consent banners with actual web QoE

Nikolas Wehner[1] · **Michael Seufert**[1] · **Raimund Schatz**[2] · **Tobias Hoßfeld**[1]

Received: 16 January 2023

Abstract

Providing sophisticated web Quality of Experience (QoE) has become paramount for web service providers and network operators alike. Due to advances in web technologies (HTML5, responsive design, etc.), traditional web QoE models focusing mainly on loading times have to be refined and improved. In this work, we relate Google's Core Web Vitals, a set of metrics for improving user experience, to the loading time aspects of web QoE, and investigate whether the Core Web Vitals and web QoE agree on the perceived experience. To this end, we first perform objective measurements in the web using Google's Lighthouse. To close the gap between metrics and experience, we complement these objective measurements with subjective assessment by performing multiple crowdsourcing QoE studies. For this purpose, we developed CWeQS, a customized framework to emulate the entire web page loading process, and ask users for their experience while controlling the Core Web Vitals, which is available to the public. To properly configure CWeQS for the planned QoE study and the crowdsourcing setup, we conduct pre-studies, in which we evaluate the importance of the loading strategy of a web page and the importance of the user task. The obtained insights allow us to conduct the desired QoE studies for each of the Core Web Vitals. Furthermore, we assess the impact of cookie consent banners, which have become ubiquitous due to regulatory demands, on the Core Web Vitals and investigate their influence on web QoE. Our results suggest that the Core Web Vitals are much less predictive for web QoE than expected and that page loading times remain the main metric and influence factor in this context. We further observe that unobtrusive and acentric cookie consent banners are preferred by end-users and that additional delays caused by interacting with consent banners in order to agree to or reject cookies should be accounted along with the actual page load time to reduce waiting times and thus to improve web QoE.

Keywords Quality of experience · Web QoE · Core Web Vitals · Crowdsourcing · Emulation · Consent banners · Cookies

Introduction

Since browsing the web is one of the most popular activities on the Internet, understanding Quality of Experience (QoE) for the web has become essential for web service providers and network operators. While currently proposed models approximate web QoE [1] either based on perceived loading times [2, 3] or on interactivity [4], no holistic approaches exist yet considering multiple potential influence factors like perceived loading time, interactivity, and visual stability.

In 2020, Google introduced the Web Vitals, a set of metrics supposed to provide guidance on how to guarantee a great user experience (UX) for web pages [5]. The Core Web Vitals (CWV) are a subset of these Web Vitals and are considered essential for every web page. The CWV consist of the largest contentful paint (LCP), the first input delay (FID), and the cumulative layout shift (CLS). The LCP is defined as

✉ Nikolas Wehner
nikolas.wehner@uni-wuerzburg.de

Michael Seufert
michael.seufert@uni-wuerzburg.de

Raimund Schatz
raimund.schatz@ait.ac.at

Tobias Hoßfeld
tobias.hossfeld@uni-wuerzburg.de

[1] University of Würzburg, Würzburg, Germany

[2] AIT Austrian Institute of Technology GmbH, Vienna, Austria

Published online: 30 June 2023

the loading time of the largest visible text or image element in the viewport, and thus, is an indicator for perceived loading time. The FID describes interactivity and is defined as the period between the first user input and the page response to said input. Finally, the CLS is an indicator for visual stability and describes the maximum layout shift of visible elements in the viewport during page load. Consequently, as the CWV cover different aspects that are also related to QoE, the CWV may have the potential to provide guidance not only for improving UX, but also for improving web QoE assessment. In this article we focus on the network-influenced aspects of web QoE and investigate the relationship between CWV and web QoE along the following research question: *To which extent do the CWV metrics correlate with the end-user's web QoE?* To answer this question, we perform both objective and subjective measurements in different Quality of Service (QoS) scenarios, which allow to understand the relationship between CWV and web QoE. The results of our work are relevant for researchers aiming to assess and quantify the QoE of interactive Web browsing as well as practitioners who want to choose the right models and metrics for optimizing the delivery of Web content.

Our objective measurements are performed using Google's Lighthouse and the top 50 Tranco web pages [6]. In particular, we analyze the sensitivity of the CWV in the network by emulating various QoS conditions. These objective measurements are complemented by QoE crowdsourcing studies, in which we emulate different LCP, FID, and CLS conditions for three custom web pages with CWeQS, our custom **C**rowdsourcing **We**b **Q**oE **S**tudy framework, which we present in detail.

Extending our previous work on this subject [7], in this article we also discuss the results of pre-studies which investigated fundamental aspects of the parmaetrization of CWeQS, as required for optimizing the design of the actual CWV crowdsourcing studies. Firstly, we determine the influence of the loading strategy of the web page elements on web QoE. In particular, we assess whether image elements or text elements should be loaded first on our custom web pages. Secondly, we investigate whether the study task of a participant directly affects web QoE, i.e., whether a different focus of participants on finding and clicking hyperlinks vs images during the study validation tasks leads to different web QoE results. This allows us to obtain better measurements by avoiding strong biases.

Addressing the relationship between the CWV and web QoE, we subsequently present the design and results of the actual CWeQS-based crowdsourcing studies for each CWV, in which participants subjectively rate the QoE as perceived after loading and interacting with the web pages. Using both kinds of measurements, we evaluate the utility of the CWV to assess web QoE. Our results suggest that the CWV seem to be less insightful for understanding and estimating web QoE than expected and, in particular, inferior to traditional metrics like Page Load Time (PLT) or Speed Index (SI).

In addition to the CWV page loading studies (and as second extension of [7]), we also investigate the influence of cookie consent banners on web QoE and the CWV. In Europe, consent banners have become ever-present due to legislation and they directly influence CWV metrics [8]: often, consent banners immediately become the largest element in the viewport, especially on mobile devices, therefore affecting the LCP. Furthermore, the first interaction on a web page is due to their omnipresence often performed with consent banners and FID is therefore determined by the responsiveness of the consent banner. Finally, consent banners often cause undesired layout shifts, directly affecting CLS. Considering three different consent banner types and loading times, we evaluate the impact of consent banners on web QoE and CWV in crowdsourcing studies with CWeQS, in particular the relevance of consent banner position, size, and loading time.

The remainder of this article is structured as follows: Sect. 2 discusses related work. The objective Lighthouse measurements and the corresponding results are presented in Sect. 3. This is followed by the description of our novel study framework CWeQS in Sect. 4. A description of our pre-studies as well as the corresponding results are described in Sect. 5. Afterwards, the CWV page loading studies are conducted in Sect. 6 and their results are discussed in Sect. 7. Finally, the CWV consent banner studies are presented in Sect. 8 and their evaluation is performed in Sect. 9, before the results are discussed in Sect. 10. Section 11 concludes the article by summarizing its key findings and implications for future work.

Related work

As regards research on the QoE of web browsing, early studies have shown that loading times are fundamental for estimating web QoE [9–11]. By now, several metrics to capture web QoE have been proposed. These web QoE metrics can be categorized into time-instant and time-integral metrics [4]. Time-instant metrics denote a specific event at which a web page load is considered complete. The most prominent time-instant metric is PLT, which specifies the point in time at which a web page is considered to be completely downloaded. Other time-instant metrics include Time to First Byte (TTFB), the point in time at which the first byte is received, Time to First Paint (TTFP), the point in time at which the first pixels are rendered, Time to Interactive (TTI), the point in time at which the web page becomes interactive, and Above the Fold (ATF), the point in time at which the content in the viewport is completely rendered. Newer metrics like ATF usually focus on the visible portions of a

web page only. This holds also for time-integral metrics. Time-integral metrics compute the integral over the complementary visual progress in the viewport by comparing mean pixel histograms over time. The best known time-integral metric is Google's Speed Index (SI). The SI quantifies how fast a web page is loaded by computing the integral of complementary visual progress based on a screen capture. Various cheap computational approximations like Byte Index (BI), progress of byte-level completion, and Object Index (OI), progress of object-level completion, have been developed and were also tested within traditional web QoE models [2]. These traditional web QoE models are usually based on the IQX and WQL hypotheses. While the IQX hypothesis assumes an exponential relationship between waiting time and web QoE [12], the WQL hypothesis assumes a logarithmic relationship on a linear ACR scale [9, 13]. Several works have shown that network quality fluctuations affect the loading process [14–16]. In addition to loading times, web QoE is also influenced by usability [17], aesthetics [18], and device type, i.e., desktop, smartphone, and tablets [1]. The authors of [19] also show that most web QoE metrics are specifically designed for desktop environments and that these metrics poorly reflect web QoE on mobile devices due to different user behavior. In this work, we focus on loading times and do not consider the impact of usability, aesthetics, and other influence factors.

In recent studies, user attention and interest have been included into web QoE assessment. Therefore, novel systems like WebGaze [20] and Eyeorg [21] have been developed. In contrast to earlier studies, user-perceived page load time (uPLT) is estimated by allowing participants to mark the point in time at which they consider a web page completely loaded. In [20], the authors show in lab and crowdsourcing studies that in contrast to uPLT both PLT and SI over- and underestimate the actual QoE severely. In [22], the authors perform crowdsourcing studies using Eyeorg to collect feedback on uPLT. They reveal that the uPLT distribution is often multi-modal, and thus requires different objective metrics for different modes.

In other works, web page contents and their influence on web QoE are assessed. In [23], the implications of failed to load web page elements are investigated. The authors show that PLT is no longer sufficient to predict web QoE and that accounting such load failures is required for improved predictions. The authors of [24] analyze not only different page loading strategies and varying page element timings on web QoE, but also the impact of the task. They reveal that QoE is highest when elements required for a task are shown early during the loading process. In general, they show that web QoE during page loading is influenced by various aspects, e.g., the page loading strategy and the total time of interaction with a web page.

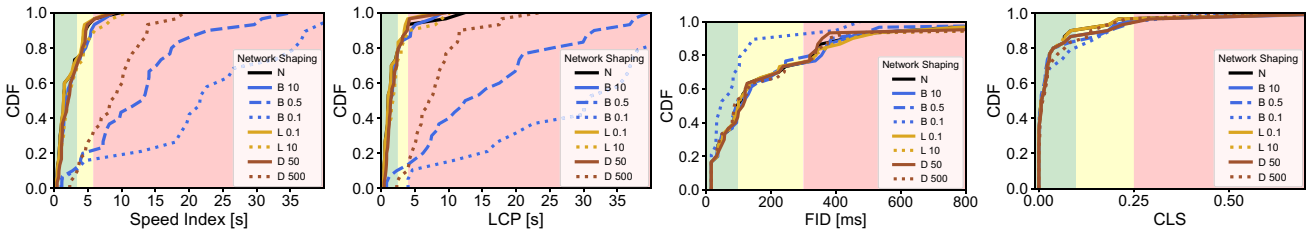
As shown in [25], the network itself and the used protocols can be considered as additional influence factors to web QoE, too. In particular, the requested protocol, the number of accessed domains, but also the location of the web page affect web QoE.

Finally, user perceptual dimensions, e.g., perceived ease of use and perceived ease of interactivity as described in [26] may also affect web QoE.

In terms of standardization activities, several recommendations have been published by ITU-T. ITU-T G.1030 [11] provides a framework to map Quality of Service (QoS) in IP networks to QoE. The recommendation ITU-T G.1031 [27] summarizes known relevant web QoE influence factors with respect to context influence factors, system influence factors, and user influence factors. A subjective testing methodology for web browsing is proposed by the ITU in recommendation P.1501 [28].

In addition to web browsing, almost any web-based multimedia service incorporates loading times in one way or the other, e.g., video streaming or music streaming. In the video streaming domain, the authors of [29] related the influence of low page load times to the subsequent video QoE. They revealed that the page load times do not affect video QoE, concluding that users expect short delays during service interaction. In the music streaming domain, the authors of [30] show that while browsing delays in a music streaming web app, e.g., for looking for a song, do not affect music streaming QoE, these delays instead degrade the QoE of the whole application.

With the widespread introduction of Internet privacy legislation, consent banners became an obligatory part of almost every website. Consent banners appear in various forms. An overview on the usage of consent banners in general is provided in [31, 32]. First performance measurements to quantify the influence of consent banners on web performance have been performed in [33, 34]. A large measurement campaign in Europe and the US has been performed in [35] to investigate the implications of consent banners. The authors show that web page performance declines after accepting privacy policies due to increased page sizes and thus slower page loading times. Most similar to our work is the work of [36], in which extensive online studies are conducted to assess the influence of common consent banner interface design decisions. Their assessment additionally includes aspects like user awareness, comprehension of choices, and privacy fatigue in users. While the work of [36] focuses on user experience, utility, and design aspects, our work emphasizes the influence of consent banner loading times and positioning on web QoE. We further want to remark that, to the best of our knowledge, the influence of consent banners on web QoE has not been investigated in literature yet.



(a) Speed Index (b) LCP (c) FID (d) CLS

Fig. 1 Distributions (CDFs) of the 75 percentile of the metrics per web page for the Lighthouse measurements where green, yellow and red areas represent Google's recommendations for good, moderate, and poor performance

While previous web QoE models usually rely on a single aspect or metric expressing the complete page loading behavior, e.g., PLT or SI, in this work we aim to model web QoE based on various aspects of a page load, i.e., loading behavior, interactivity, and visual stability, as defined by the CWV. Consent banners also exert impact on the CWV and must therefore be additionally considered [8]. In previous work, however, consent banners have only been investigated via passive measurements without any user involvement, while we perform dedicated subjective measurements in this work.

Objective measurements with Lighthouse

Study setup

In the following, we conduct and evaluate measurements using Google's Lighthouse,[1] which is a tool for improving the quality of web pages and is able to run a variety of tests against a web page while monitoring various performance metrics like the CWV and SI. We perform these measurements for two reasons. First, we are able to observe the potential range of CWV scores in the wild, which allows us to validate Google's recommendations on the one hand, and which provides us guidance on how to determine the conditions of the subjective studies on the other hand. Second, we are able to quantify the influence of QoS on the CWV, which may be beneficial for estimating the CWV from network measurements later.

Our Lighthouse study setup is a dockerized environment, in which we perform headless Lighthouse runs with NodeJS and use Linux tc to emulate varying network conditions on the network interface, on which Lighthouse runs. For the purpose of emulation, we use docker-tc provided on GitHub.[2] All Lighthouse reports are then stored in a MinIO[3] instance.

The utilized network shapings include adding one-way delay to the packet transmissions (50, 100, 250, and 500ms), introducing different packet losses (0.1, 1, and 10%), and limiting the available bandwidth (0.1, 0.5, 1, and 10 Mbps). We performed at least 30 runs for all top 50 Tranco web pages [6] for an emulated mobile device and an emulated desktop device. Our evaluation revealed that mobile and desktop measurements behaved similar except for increased PLTs on desktop and increased CLS values on mobile.

Influence of QoS on CWV and SI

Figure 1 therefore depicts only the results for selected network shaping conditions of the desktop Lighthouse measurements in form of CDFs. As Google recommends that 75% of web page visits should provide a good experience[5], for each web page, we consider the 75 percentile of LCP, FID, CLS, and SI over all measurement runs for this web page. The CDFs depict the distribution of these 75 percentiles over all 50 web pages. The CDFs are styled and labeled according to their network shaping conditions, whereby D denotes packet delays, L denotes packet losses, B denotes bandwidth limitations, and N denotes no shaping. Additionally, the green, yellow, and red areas represent Google's recommendations for good, moderate, and poor performance. In general, the CDFs indicate that most pages show good and moderate performance.

Considering the three CWV, it can be observed that only LCP shows a substantially different behavior when facing different network conditions (Fig. 1b). In particular, the LCP behaves very similar to the SI (cf. Fig. 1b and a), and both easily end up with poor performance as soon as the network conditions are really bad. This is reasonable as both metrics represent loading behavior, and thus, indicates that the LCP may act as a proxy for the SI. In contrast, FID and CLS are

[1] https://developers.google.com/web/tools/lighthouse.

[2] https://github.com/lukaszlach/docker-tc.

[3] https://min.io/.

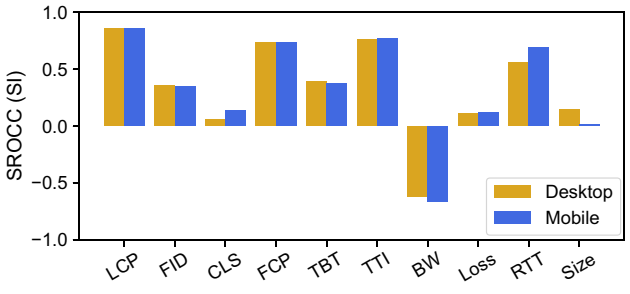


Fig. 2 Correlation (Spearman) of (Speed Index) SI with the CWV and other metrics used for Lighthouse measurements

barely influenced by deteriorating network conditions (cf. Fig. 1c and d). This suggests that FID and CLS strongly depend on the design of the individual web pages and that in-network monitoring of these metrics proves to be difficult.

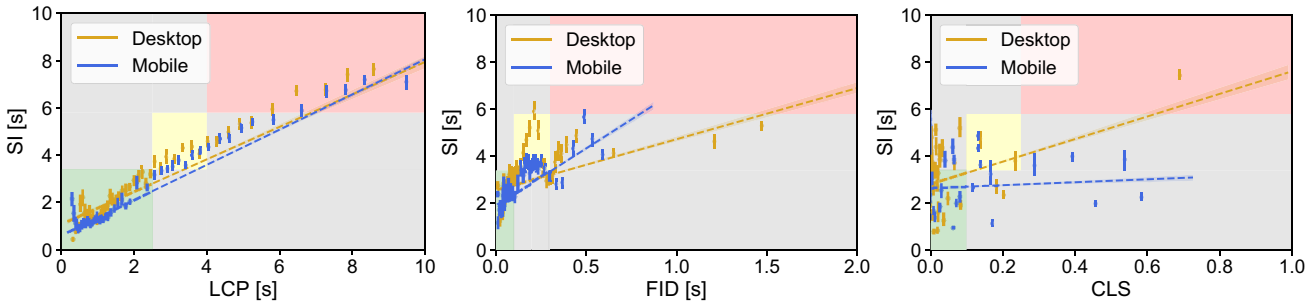
Summarizing, we observed that most web pages align with Google's recommendations and that LCP is the only CWV metric affected by the network. Moreover, by measuring popular web pages in the wild, we identified meaningful study conditions for the crowdsourcing QoE studies.

Correlations of CWV and SI

Last but not least, we consider the correlations, in particular the Spearman Rank Order Correlation Coefficient (SROCC) between the CWV and the SI across the desktop and mobile measurements. Figure 2 depicts those correlations, whereby the x-axis presents the CWV and other web page metrics, while the y-axis corresponds to the SROCC, computed with respect to the SI. In addition to the CWV, the x-axis denotes the first contentful paint (FCP), the total blocking time (TBT), the time to interactive (TTI), the bandwidth, the packet loss, the round trip time in the network (RTT), and the document size in bytes. TBT is another Google metric, which corresponds to the time between FCP and TTI [37]. The correlations for the desktop measurements are provided in gold, while the correlations for the mobile measurements are provided in blue.

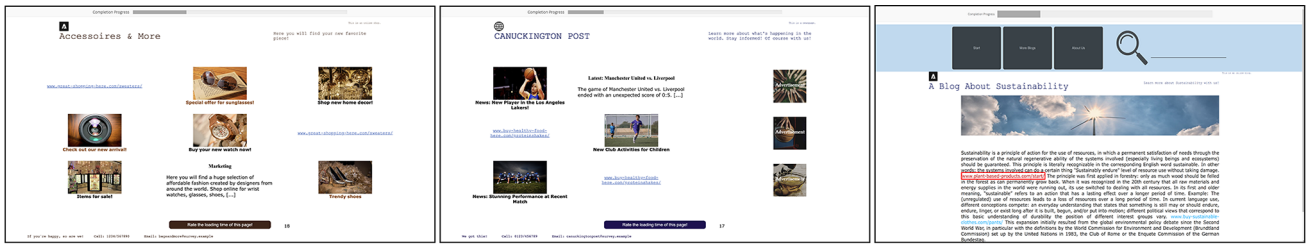
First of all, we can observe that all metrics except bandwidth are positively correlated to the SI and that for the majority of metrics the SROCC for desktop and mobile measurements are similar. In particular, LCP (0.82), FCP (0.69), and TTI (0.73) are strongly correlated to the SI. Since LCP, FCP, and TTI are closely related, this comes as expected. Similar to Fig. 1, we observe that FID is only moderately correlated and that CLS shows no correlation at all. When investigating the network aspects bandwidth, packet loss, and RTT, we observe high negative correlations for bandwidth (desktop: −0.62, mobile: −0.66), low positive correlations for packet loss (desktop: 0.11, mobile: 0.12), and moderate positive correlations for RTT (desktop: 0.54, mobile: 0.68) and for latency (desktop: 0.55, mobile: 0.59). These findings come again as expected since network performance is a critical aspect for web performance. The low correlation between packet loss and SI may be attributed to the fact that the bandwidth was still sufficient to load the web page quickly despite packet losses. Highly interesting here, is the fact that the document size in bytes is barely correlated to the SI, in particular for the mobile measurements the SROCC is close to 0.

Both Figs. 1 and 2 indicated a monotone relationship between LCP and SI and no relationships between FID and SI and CLS and SI. We therefore investigate once again their relationships by comparing the metrics directly in Fig. 3. The x-axis denotes the CWV metric and the y-axis the SI. Green, yellow, and red areas denote good, moderate, and poor performance for both CWV and SI as recommended by Google and grey areas denote "undefined" areas, i.e., areas where either the performance of CWV and/or SI are incompatible. Again, the measurements are separated by desktop measurements (gold) and mobile measurements (blue). Figure 3a shows that LCP and SI result in an almost perfectly linear relationship (indicated by the dashed regression lines), which is well aligned with the Google recommendations,



(a) Largest Contentful Paint (b) First Input Delay (c) Cumulative Layout Shift

Fig. 3 SI versus the three different Core Web Vitals (LCP, FID, CLS) as measured on different platforms using Lighthouse where green, yellow and red areas represent Google's recommendations for good, moderate, and poor performance



(a) Shopping page (b) News page (c) Blog page (with CLS)

Fig. 4 Custom web page types implemented in CWeQS

and that the relationship is very similar between desktop and mobile measurements. This verifies that LCP may definitively act as a good proxy for the SI. For FID, in contrast, desktop and mobile measurements differ strongly and there is no obvious relationship between FID and SI visible (cf. Figure 3b). For CLS, it is even worse since Fig. 3c shows only a random pattern. Thus, in both Fig. 3b and c, the fitted regression lines do not convey meaningful information.

This confirms our previous hypotheses that LCP acts as a proxy for the SI, while FID and CLS strongly depend on other factors, e.g., the used end-device or the web page design.

CWeQS: crowdsourcing web QoE studies

For our web QoE studies we rely on CWeQS,[4] a custom crowdsourcing framework, which allows to fully control the loading behavior of custom web pages. Moreover, it provides a rich set of required features for crowdsourcing QoE studies, such as questionnaires, preparation of study conditions, and means to assess reliable study execution. It is based on JSPsych [38], a JavaScript framework for browser-based studies.

To exert complete control over the web page loading behavior, CWeQS follows a top-down approach, i.e., instead of varying network conditions to generate a variety of web page loading behaviors, we emulate these behaviors independent of the network conditions by manipulating the appearance of the DOM elements with arbitrary timings.

In detail, these timings are realized with the *setTimeout()* functionality of JavaScript, which executes an arbitrary function after a specified timeout has been reached. Here, this function corresponds to the rendering of an element, i.e., setting the element's visibility in CSS to true, and the timeout corresponds to a specified loading time. Each page element is thus assigned a loading time or timeout, respectively, and *setTimeout()* is called simultaneously on all page elements as soon as the participant triggers the page load. Page elements are then rendered as soon as their timeouts have expired.

In total, we use four parameters which specify a complete page load. These four parameters, named FP (first paint of small header elements), TTText (time to first substantial text), TTImage (time to first substantial image), and PLT (page load time), are sorted in ascending order, i.e., $FP \leq TTText \leq TTImage \leq PLT$. Note that PLT here corresponds to the ATF time, as we only show elements in the viewport. These parameters are evenly spaced with respect to the PLT, and to avoid an unrealistic step by step loading behavior, where many elements appear at once, we additionally use a $\beta(7.2, 0.8)$-distribution to smooth the loading process for around half of the elements. As a consequence, the mean loading time of the distributed elements is 90% of the actual specified loading time with a standard deviation of 10%.

As this approach requires us to know beforehand which integral elements appear on a web page, we can only use custom web pages in CWeQS at the moment. To rule out any negative network impacts during the study, we preload all web page elements on client-side with a JSPsych plugin when the framework is first loaded in the browser. We implemented three custom pages, depicted in Fig. 4, which represent common web page categories, namely, an online shop and a news page, consisting of a mix of texts and images, and a blog page, consisting of much text and a single large picture.

To align with best practices for crowdsourced QoE studies [39], CWeQS requires a method for evaluating the validity of participants. Our framework provides two different types of validation: image validation and hyperlink validation. With both types, participants have to interact with the web pages, which provides the additional benefit of making the study tasks more realistic. With image validation, participants are primed in the instructions to mark target images on a web page by clicking them. A random number (up to three) of these target images are inserted in the web page by randomly exchanging the actual images with the target

[4] https://github.com/lsinfo3/CWeQS.

images. Any image on the page that is clicked is then framed with a red border. The number of total target images as well as correctly identified target images are then used to identify unreliable study participants, which are excluded before the evaluation. Hyperlink validation works the same way except that hyperlinks, i.e., pieces of highlighted text, are supposed to be clicked by participants. Both a marked and a not yet marked hyperlink are illustrated in Fig. 4c.

Finally, CWeQS can be operated in two different execution modes: *study mode* and *standalone mode*.

Study mode

The procedure in *study mode* consists of seven phases: First, during study startup, a chain of checks is performed whether a user is allowed to participate in the study. This includes, for example, the verification of the participant ID and browser size requirements. After providing a first set of instructions, participants are asked for demographic information and browsing habits. This is followed by instructions, in which the actual study procedure is explained and in which participants are briefed what they are supposed to do. Then, training stimuli are shown to the participants to prime them on the task. This is again followed by another set of instructions, before the actual test stimuli are shown to the participants. Participants are asked for their opinion immediately after each stimulus. After observing all test stimuli, participants are rewarded with a verification code. A training or test stimulus hereby consists of the emulated page load and the subsequent questionnaire, in which participants rate the perceived loading time on the Absolute Category Rating scale [40].

Standalone mode

To perform a single page load in *standalone mode*, only the loading parameters of each element have to be passed via the URL. With these URL parameters, we are then able to populate CWeQS with the required configuration and start the timings of a page as usual. This mode allows us to *replay* the stimuli observed by the participants during the study locally in order to compute additional metrics. Using this method, we additionally compute the SI of all stimuli in this work. This is achieved by performing screen captures while replaying the logged configurations and then computing the SI with existing scripts provided by WebPageTest[5] based on these screen captures. We automate this task with Selenium[6] and FFmpeg.[7]

[5] https://github.com/WPO-Foundation/visualmetrics.
[6] https://www.selenium.dev/.
[7] https://www.ffmpeg.org/.

Pre-studies using CWeQS

First of all, we conduct pre-studies, in which we evaluate the influence of the loading strategy and the influence of the validation task when using CWeQS. It is important to understand and quantify these two phenomena in order to avoid any unwanted influences on the results of the planned web QoE crowdsourcing studies. For the loading strategy, we investigate whether the order of the loading times of images and text elements in a web page influence the user ratings. For the validation task, we compare hyperlink validation with image validation as described in the previous section.

For these pre-studies, we use CWeQS in its most generic configuration without emulating specific metrics, i.e., the custom web pages are simply loaded according to the configured element timings.

For the investigation of the loading strategy, we consider a full factorial design of our four parameters FP, TTText, TTImage, and PLT using a step size of either 0.5 s or 2.5 s between the parameters. This results in the sixteen conditions presented in Fig. 5. We conduct two studies, whereby in the first study the text elements are shown first and in the second study the image elements are shown first. This means, we simply swap the TTText and TTImage values. These conditions apply for each of our three custom web pages. For both studies, participants have to perform image validation.

For the analysis of the validation task, we select only a subset consisting of four conditions of all sixteen conditions. This subset consists of the fastest loading strategy, the slowest loading strategy, and two loading strategies in between. Here, we select a subset only to save time since we expect that the validation tasks result in marginal differences only.

The pre-studies were conducted in March 2021 with the crowdsourcing platform Microworkers.[8] Participants were shown six to eight randomly selected test stimuli in total. The selected types of pages were uniformly distributed. After each stimulus, participants answered a single question *How did you experience the loading of the last page?* on a classical five-point ACR scale.

We excluded participants if they marked less than 80% of the displayed hyperlinks or images correctly and we excluded participants giving the same rating for each test stimuli, even though the stimuli differed strongly. After this very strict filtering, 92 unique participants and 951 rated test stimuli remained after filtering for the loading strategy investigation, while 137 unique participants and 979 test stimuli ratings remained for the task investigation.

[8] https://www.microworkers.com/.

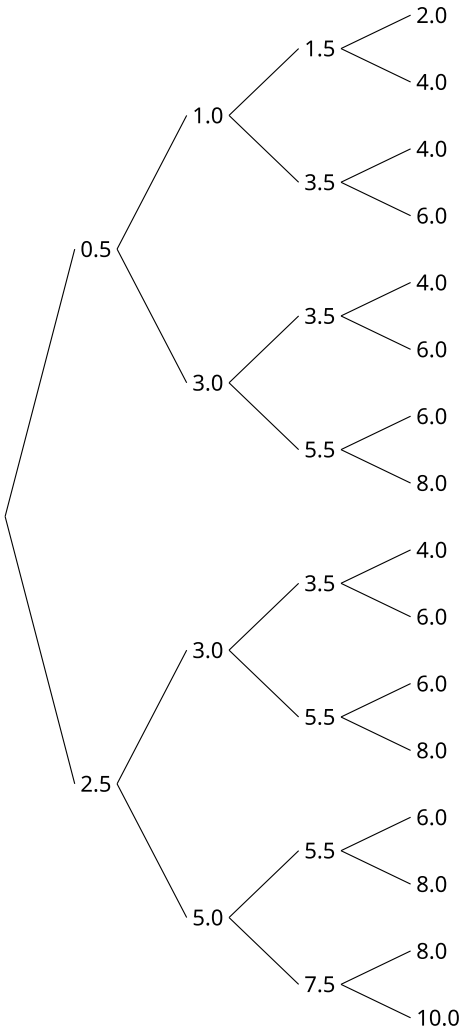


Fig. 5 Pre-study conditions covering different combinations of loading times (in seconds) where values from left to right denote FP, TTText/TTImage, TTImage/TTText, and PLT

Influence of loading strategy

First, we analyze the influence of loading strategy, i.e., whether there is a difference when image elements appear before text elements or vice versa. We therefore depict the results of our pre-studies in Fig. 6, in which the MOS is provided along with the 95% confidence intervals on the y-axis. Each bar is colored depending on the loading strategy (text elements first in blue, image elements first in gold). The bars show that no significant differences between the loading strategies can be observed. This is also confirmed when performing condition-wise Mann-Whitney-U tests between the two groups. It is further clearly visible that conditions with a low PLT (2–4 s) result in a significantly higher MOS,

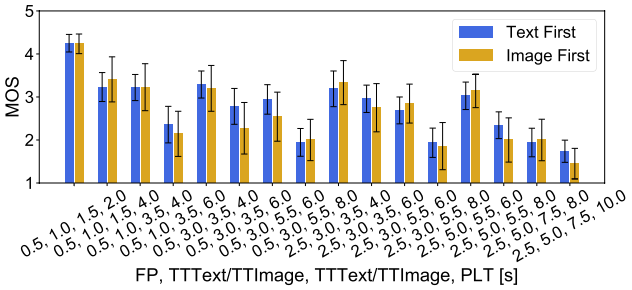


Fig. 6 Web QoE (MOS) for different loading strategies (bar colours) and text/image positionings where each study condition is denoted as tuple of FP, TTText/TTImage, TTImage/TTText, PLT (in seconds)

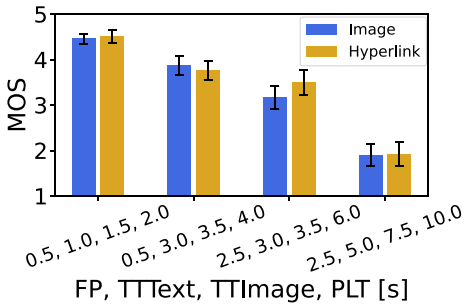


Fig. 7 Web QoE (MOS) for the validation task for selected loading time conditions

independent of the loading strategy. Our results therefore indicate that the order of elements is not relevant here compared to influencing factors like PLT.

Influence of validation task

Next, we investigate whether the crowdsourcing task itself, i.e., hyperlink validation or image validation, affects the ratings of the participants. We therefore consider the selected conditions in Fig. 7, in which again the MOS along with the 95% confidence intervals is provided. This time, bars colored in blue correspond to image validation, while golden bars correspond to hyperlink validation. Here, we observe that the task does not affect the MOS significantly for any of the four selected conditions. A Mann-Whitney U test confirms this notion. As a consequence, we observe no impact of the validation task on the MOS.

For the remaining studies of this article, we use hyperlink validation since no differences between the validation methods were found and hyperlinks are more unobtrusive. This unobtrusiveness may be beneficial when investigating the CWV later to be able to highlight occurring effects better.

Table 1 CWV page loading study conditions

CWV	Parameter values	PLT [s]
LCP [s]	(1.00, 1.50, 2.00)	2.0
	(1.00, 1.50, 2.50, 3.75, 5.00)	5.0
	(1.00, 1.50, 5.00, 7.50, 10.00)	10.0
FID [s]	(0.1, 0.3, 0.5, 1.0, 2.0)	2.0
	(0.1, 0.3, 0.5, 1.0, 2.0)	5.0
	(0.1, 0.3, 0.5, 1.0, 2.0)	10.0
CLS	(0.0, 0.1, 0.2, 0.3) × (PLT/2, PLT)	2.0
	(0.0, 0.1, 0.2, 0.3) × (PLT/2, PLT)	5.0
	(0.0, 0.1, 0.2, 0.3) × (PLT/2, PLT)	10.0

CWV page loading studies

We conduct a QoE study for each of the three CWV with CWeQS. For this, we select a subset of realistic parameter values from the parameter ranges observed in the Lighthouse measurements. In particular, we test three different PLTs (2, 5, and 10 s) in each study and test no more than five manifestations of each CWV. A comprehensive overview on all crowdsourcing study conditions and CWV parameters is given in Table 1. In the following, the realization of the CWV metrics in CWeQS is outlined.

Largest contentful paint

We simulate LCP by randomly selecting one of the available images on a web page and by increasing width and height of this image significantly to fixed values. Width and height of the LCP are not varied throughout the study. This enlarged image is then rendered as usual to a specified time. We design these rendering times in dependency of the PLT. In detail, we use 50%, 75%, and 100% of the PLT as time for displaying the LCP. For PLTs of 5 and 10 s, we additionally use LCPs of 1 and 1.5 s to be able to compare LCP across the different PLTs.

First input delay

To simulate FID, we monitor the user interactions with a web page and artificially delay the web page response to the first user interaction, i.e., a click to an image or a hyperlink, by again utilizing the *setTimeout()* functionality of JavaScript. All additional user interactions occurring after the first interaction and during the FID are blocked and queued. All user interactions are then responded to, i.e., by marking the clicked element with a red box, simultaneously as soon as the FID timeout has passed.

Note that FID is triggered by the user's first click on a visible interactive event, which can happen at any time (even after the PLT). However, participants are supposed to experience the FID during the page load, as it would be unnatural to have an input delay after the page is completely loaded. Thus, we additionally instructed the participants to click the targets as fast as possible after their appearance.

The selected FID values are partly recommended by Google, and partly determined in dependency of a PLT of 2 s.

Cumulative layout shift

CLS represents the largest observed layout shift score during the entire page load. A page load can hence contain multiple layout shifts. Layout shift scores are computed by multiplying the *impact fraction* with the *distance fraction*. The *impact fraction* defines the fractional area of the viewport, in which unstable elements have moved between two frames. If the viewport is already filled completely during a layout shift, the impact fraction is 1. The *distance fraction* defines the largest fractional distance any of the unstable elements has moved in the viewport.

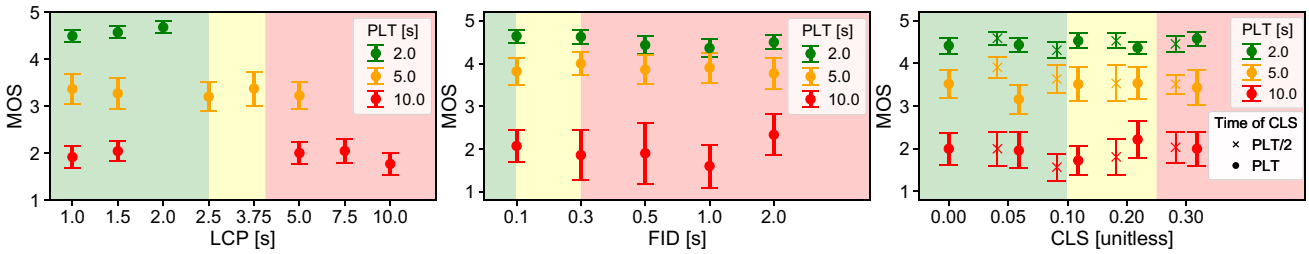
To simplify the emulation of CLS, which depends on basically all elements in the viewport, we perform layout shifts by displaying a banner with a specific height on top of the original page at a specific time. An example can be seen in Fig. 4c, where the CLS is caused by displaying the blue banner above the actual page content, shifting all other elements, including headline, image, and text, towards the bottom. We consider two points in time for performing the layout shift. In the first case, we perform the layout shift at the end of the page load. Since the whole viewport is occupied then, the *impact fraction* is automatically 1. Consequently, the *distance fraction*, i.e., the banner height relative to the viewport size, fully determines the CLS score. In the second case, we perform the layout shift at half of the PLT, at which time only the first row of elements and the header are in the viewport. This gives a fixed *impact fraction*, which can now be multiplied with the *distance fraction* from above to obtain the desired CLS score.

In our CLS study, we provide stimuli of both use cases to the participants and use the CLS values provided in Table 1, which are again aligned to the recommendations of Google.

Results of CWV page loading studies

All three studies were conducted in December 2021 and January 2022 using the crowdsourcing platform Microworkers.[9] All Microworkers meeting the hardware requirements, e.g., a minimum screen size, were allowed to participate and

[9] https://www.microworkers.com/.



(a) LargestContentfulPaint (b) FirstInputDelay (c) CumulativeLayoutShift

Fig. 8 Web QoE (MOS) for different settings of Google's Core Web Vitals (LCP, FID, CLS) and PLTs in page loading studies

were rewarded with 0.25 U.S. dollar. After ensuring that the browser size was large enough to fully display the page, participants were shown six randomly selected test stimuli in total. The selected types of pages, i.e., news, shopping, and blog page, used for these stimuli were also uniformly distributed. After each stimuli, participants answered a single question *How did you experience the loading of the last page?* on a five-point ACR scale ranging from bad to excellent.

512 participants completed the LCP study, while the FID study had 227 participants and the CLS study had 417 participants. We excluded participants if they marked less than 80% of the displayed hyperlinks correctly. For the FID study, we additionally removed participants who took longer than five seconds to perform the first click after the first hyperlink was rendered. Finally, we excluded participants giving the same rating for each test stimuli, even though the stimuli differed strongly. After this very strict filtering, 183 participants remained for the LCP study, which rated a total of 1098 test stimuli. For the FID study, 140 participants and 840 rated test stimuli remained, and for the CLS study, 207 participants and 1014 rated test stimuli remained after filtering.

All valid 323 participants were older than 18 years. 33.9% were women, 65.5% were men, and the rest were diverse. 54.7% of participants were from Asia, followed by 18.6% and 17.5% from Europa and South America, respectively. More than 93% of the participants use the Internet daily.

Relation of CWV to web QoE

Figure 8 shows the mean opinion score (MOS) along with the 95% confidence intervals for each crowdsourcing study in dependency of the PLT. The x-axis describes the CWV conditions, while the y-axis denotes the MOS. The different colors of the bars illustrate the total PLT. As we tested two different event times for CLS, we added an additional legend in Fig. 8c, which states the time of the layout shift in dependency of the PLT. Thin bars correspond to PLT/2, while regular bars correspond to PLT.

In all three figures, it can be observed that PLT is the main influence factor, as indicated by the different MOS regions around 4.5 for a PLT of 2 s (green), around 3.5 for 5 s (yellow), and around 2 for 10 s (red). What is highly surprising is that these MOS regions are stable with respect to the CWV conditions. This means that, considering the same PLT value, no variation of the LCP, FID, or CLS parameters has a significant impact on the MOS. This also holds when considering Google's recommended parameter ranges for good, moderate, and poor performance highlighted by the green, yellow, and red areas. Thus, these results indicate that the CWV metrics do not properly express the actual web QoE in terms of MOS.

Impact of PLT and SI on web QoE

As our crowdsourcing studies found no significant impact of CWV on web QoE, we will now investigate the impact of PLT and SI in more detail. We use the standalone mode of CWeQS and compute the SI observed by participants during the studies by replaying logged configurations locally. Figure 9 depicts the SROCC for the CWV, the PLT, the SI, and the MOS. Both PLT (−0.74) and SI (−0.70) show a

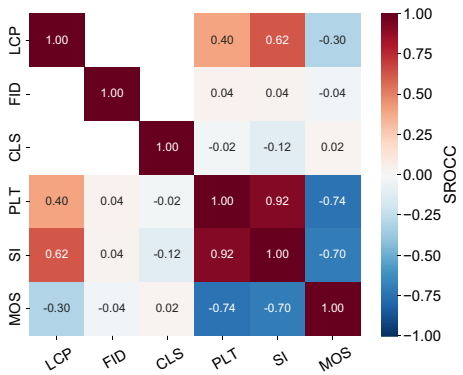


Fig. 9 Correlation between subjective and objective metrics used in CWV page loading studies

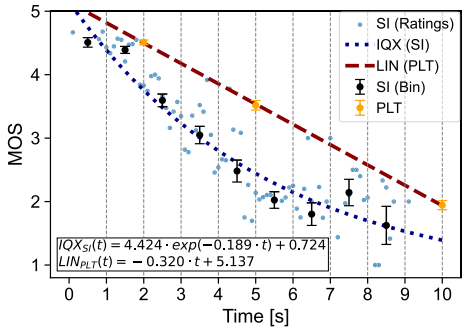


Fig. 10 Relation of PLT and SI to MOS in page loading studies

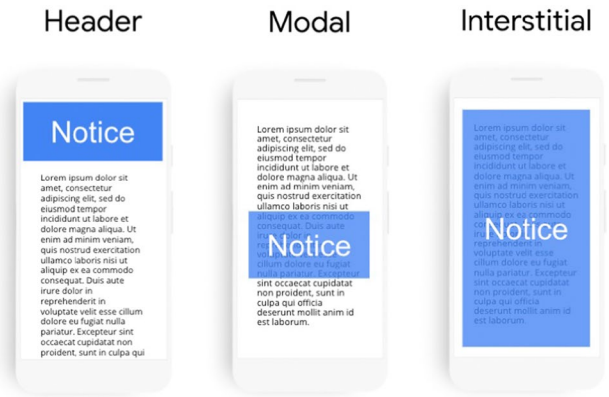


Fig. 11 Considered banner types of [8]

high negative SROCC to the user ratings, which comes as expected considering results of previous QoE studies [2, 10].

Figure 10 visualizes the relationship between averaged user ratings and SI as light blue dots on the continuous SI scale. We also bin the SI in 1 s intervals, and visualize the MOS along with the 95% confidence intervals for each bin in black. When fitting the MOS values for every bin, we observe that both IQX [12] and WQL hypothesis [9, 13] clearly apply for SI, which confirms the results of [2]. The best fit, slightly better than WQL, is $IQX_{SI}(t) = 4.424 \cdot \exp(-0.189 \cdot t) + 0.724$, which gives a very high coefficient of determination $R^2 = 0.9544$. Comparing this fit with the previous work of [2], we see that our model shows a steeper slope and uses almost the full range of MOS in the considered SI range, which indicates that the participants in our studies are less tolerant with respect to the page loading times as expressed by SI.

Still, when visualizing the MOS and 95% confidence intervals for all three investigated PLT values, as depicted in yellow, we clearly see a linear trend. This is confirmed by an almost perfect fit $LIN_{PLT}(t) = -0.320 \cdot t + 5.137$ with $R^2 = 0.9998$. This is a surprising finding considering that previous web QoE studies did not find linear relationships between PLT and MOS. When comparing our PLT model to the PLT models of [2] and [13], our model is more tolerant with respect to short waiting times. However, the covered MOS range of our model is higher than in [2] and more similar to the PLT models of [13].

CWV consent banner studies

Finally, we use CWeQS to conduct crowdsourcing studies in which we investigate the influence of consent banners on web QoE. Consent banners are closely related to the CWV since they often cause layout shifts (CLS), become the largest element in the viewport (LCP), have to be accepted first (FID), and sometimes even hide relevant content. Due to legislation in Europe, consent banners are nowadays omnipresent and a conscious usage with respect to web QoE is therefore crucial.

Table 2 Conditions for CWV consent banner studies

Banner type	Banner time [s]	PLT [s]
Header, Modal, Interstitial	0.50, 1.00, 2.00	2.00
	1.25, 2.50, 5.00	5.00
	2.50, 5.00, 10.0	10.0

Consent banners come in many different forms and may appear at various points in time during a page load, even though it is common that they are usually loaded in the beginning of the page loading and rendering cycle. Potential factors that may affect web QoE are therefore position and size of consent banners as well as their loading time. For example, consent banners which are loaded late can prolong the experienced page load process in a negative way or hide to the user that the page load is still going on in a positive way. Google summarizes their best practices for consent banners with respect to performance and UX in [8].

Being guided by Google's best practices, we consider three of the proposed consent banner types in our crowdsourcing studies. In detail, we focus on *header*, *modal*, and *interstitial* as shown in Fig. 11. We chose these three banner types since they strongly differ in position and size and since they are the most common types observed in the field by ourselves. We emulate these banner types in CWeQS by simply overlaying an additional element in the case of *modal* and *interstitial* and by adding a *header* element on top of the page, identical to what we did when emulating CLS. The *header* element hereby has a height of 20% of the actual viewport height. The *interstitial* element covers the viewport fully without showing any website contents and the *modal* element has a width and height of 50% of the viewport width and height, respectively, and is centered in the viewport. With respect to the loading times of the

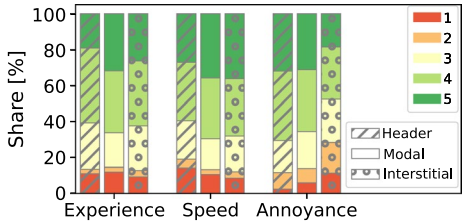


Fig. 12 Distributions of user ratings (1 = bad, 5 = excellent) of different QoE dimensions for different banner types

consent banners, we consider three different points in time relative to the PLT: $0.25 \cdot PLT$, $0.5 \cdot PLT$, and $1.0 \cdot PLT$. In a full-factorial design, this leads us to 27 study conditions as indicated in Table 2.

We again adopt the study and stimuli design of CWeQS, but add the obligatory additional task of accepting or rejecting the loaded consent banner, before being able to perform the validation task, for each stimulus.

Results of CWV consent banner studies

The CWV consent banner studies were conducted in June and July 2022, again using CWeQS and the crowdsourcing platform Microworkers[10] for recruiting and managing participants. Again, all Microworkers meeting the hardware requirements, e.g., a minimum screen size, were allowed to participate and were rewarded with 0.25 U.S. dollar. After the common checks, participants were shown five randomly selected test stimuli in total. The selected types of pages were uniformly distributed. After each stimulus, participants answered three questions:

- How was your overall experience of the last page?
- How did you experience the loading of the last page in terms of speed?
- How annoying was the consent banner of the last page?

Both the overall experience and loading speed question are answered on the classical five-point ACR scale, ranging from bad to excellent, while the annoyance question is answered on a five-point DCR scale, ranging from extremely annoying to not annoying at all.

We excluded participants if they marked less than 80% of the displayed hyperlinks correctly or if they took longer than five seconds to click the consent banner after its loading. Finally, we excluded participants giving the same speed rating for each test stimuli, even though the stimuli differed strongly. After this very strict filtering, 87 participants and

[10] https://www.microworkers.com/.

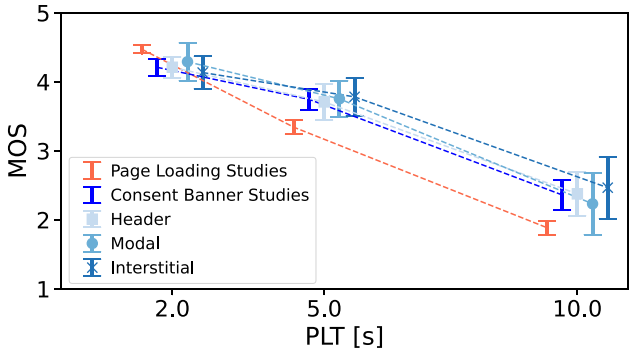


Fig. 13 QoE (MOS in terms of overall experience) at different page load times for CWV page loading and CWV consent banner studies and additionally, split by banner type

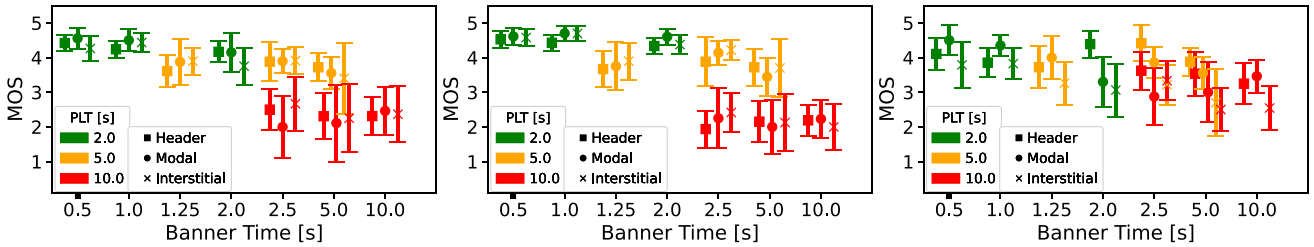
455 rated test stimuli remained after filtering, so 17 ratings on average for the 27 conditions and around 100 ratings per PLT. The demographic information in these studies is very similar to the one stated above for the CWV page loading studies.

Relationship between consent banner type and MOS

First, we compare the three different banner types across the different ratings. Figure 12 depicts the distributions of ratings for each banner type. The x-axis denotes the considered rating of experience, speed, or annoyance, and the y-axis corresponds to the share of ratings in percent. Hatched bars correspond to the consent banner type *header*, plain bars to *modal*, and dotted bars to *interstitial*. The stacked bars are colored according to ratings, whereby 1 is bad and extremely annoying and 5 is excellent and not annoying at all.

The figure reveals that the ratings for speed and experience are highly similar for all three banner types. More interesting is the fact that around half of the participants experienced the banner type *interstitial* as extremely annoying to annoying (1–3). In contrast, only 25% experienced *header* as extremely annoying to annoying (1–3), while *modal* is in between *header* and *interstitial*. Since experience does not seem to be affected by any banner type, we thus conclude that *header* should be chosen as banner type as it reduces the remaining factor annoyance strongly. This result is not completely surprising since *header* is the most unobtrusive banner type of the three due to its limited size and marginal position here.

Additionally, Fig. 13 compares the MOS of the CWV page loading studies, the CWV consent banner studies, and each consent banner type, grouped by PLT and along with the 95% confidence intervals. The dotted lines represent the trend for each group. Note that the bars for the page loading studies correspond to the values for PLT depicted in Fig. 10. The figure shows that all banner types behave similarly and



(a) Overall experience (b) Loading speed QoE (c) Banner annoyance

Fig. 14 MOS for different QoE dimensions as influenced by consent banner time and banner type

that the MOS of the CWV page loading studies studies and the MOS of the CWV consent banner studies are slightly different, in particular for higher PLTs. Only for a PLT of 2 s, the MOS of the CWV page loading studies is slightly higher than the MOS of the CWV consent banner studies. For PLTs of 5 and 10 s, participants in the CWV consent banner studies seemed to be more tolerant with respect to long loading times as the MOS is here slightly higher. This more tolerant behavior can be likely ascribed to the obfuscation of the actual PLT due to distracting consent banners. Therefore, we conclude that consent banners affect web QoE in general, but only marginal.

Influence of consent banner time on MOS

Next, we investigate the influence of the loading time of a consent banner on the MOS. The results for each rating are illustrated in Fig. 14. The x-axis denotes the time when the consent banner was displayed and the y-axis denotes the MOS. Each error bar shows the MOS along with the 95% confidence intervals per banner time, banner type, and PLT. Error bars are colored according to the PLT of each condition and the markers of the error bars change for each banner type (*header*: square, *modal*: circle, *interstitial*: cross). When observing the results for the experience rating and the speed rating in Fig. 14a and Fig. 14b, respectively, it is visible that again the PLT has the highest impact on the MOS and that experience rating and speed rating are quite similar. We can see that the MOS is stable across the different consent banner loading times and that it only depends on PLT. Also, no consent banner type is outperforming others in terms of MOS. When relating the consent banner loading time and the annoyance rating, we note that the annoyance differs between the individual consent banner types slightly and that some trends are visible. In particular, the *header* consent banner seems to be the least annoying banner type when participants are confronted with higher PLTs and higher banner times in general. In contrast, with a banner time of 0.5 s or 1.0 s and a PLT of 2.0 s, *header* is no longer considered better than the other banner types, but instead *modal* seems to be preferred. Even though the question about annoyance has not been tailored towards loading speed, there is also a weak negative linear relationship between MOS and PLT visible. This indicates that high PLTs also affect participants subconsciously. Summarizing, we note that banners of type *header* should likely be used to optimize QoE and that the point of time when a consent banner appears does not affect experience, but annoyance only.

Analysis of clicking behavior

In the following, we analyze the clicking behavior of the participants for the consent banner studies. In particular, we analyze the reaction times, i.e., first clicks on banners, of participants to the varying banner types and we quantify the additional delay until first content interaction caused by consent banners. Let us denote BT as the loading time of the consent banner and BC as the point in time when a participant clicks on a banner. The metric BC-BT thus quantifies the reaction time of a user after the consent banner has loaded. The relation between PLT and BC and BT is denoted by PLT-BC and PLT-BT, which denote the remaining time of the PLT after consent banner click or after consent banner loading time. Note that PLT-BC might become negative if a user clicks on the consent banner after the PLT has passed.

First, we analyze the distribution of the reaction times for each banner type in Fig. 15. The figure shows the CDFs for

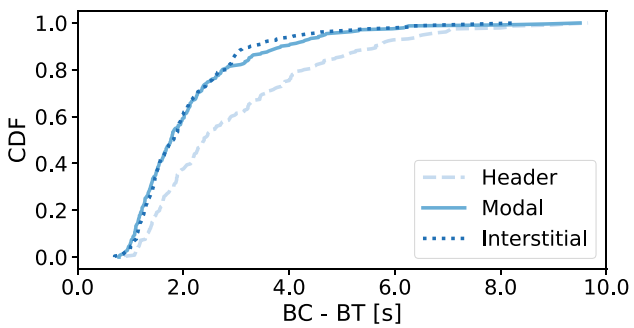


Fig. 15 Distributions (CDF) of reaction time for each banner type

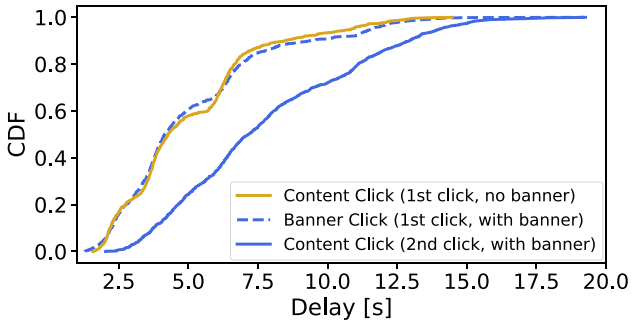


Fig. 16 Distributions of content and banner click delays for conditions with (blue) and without (yellow) consent banner

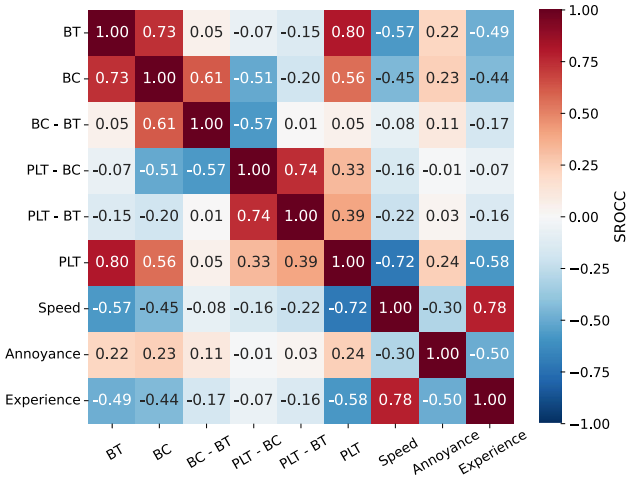


Fig. 17 Correlation (Spearman) between objective and subjective metrics used in the CWV consent banner studies

the reaction time (BC-BT) grouped by banner type. We can observe that reaction times for *modal* (mean: 2.17, std: 1.34, median: 1.77) and *interstitial* (mean: 2.10, std: 1.17, median: 1.82) are very similar, while the CDF for *header* deviates significantly. Here, the reaction times are in general higher with a mean of 3.02 s and a median of 2.39 s. This can be likely attributed to the positioning of the banner and buttons, respectively. While the buttons for the *modal* and *interstitial* banner are centered, the buttons for the *header* banner are in the upper left of the viewport, thus taking more time to be located and clicked.

Secondly, we quantify the additional delay caused by consent banners. For this purpose, we compare in Fig. 16 the distributions of the first content clicks of the FID study from Sect. 6, where no consent banners had been shown, with the distributions of the first consent banner clicks and the first content clicks. A content click hereby describes the first interaction with an arbitrary web page element (except for the consent banner), e.g., a hyperlink or an image. Note that in the consent banner studies, the first content click actually is the user's second click as the consent banner has to be clicked first. The CDFs denote the passed time between the start of the page load and the corresponding kind of interaction. We can easily see that the first clicks on a web page (content element for FID study in golden and consent banner for consent banner studies in dashed blue) are similarly distributed. The results therefore suggest that users in our study interact with consent banners in the same way as with actual content. When comparing the first content click of the FID study (golden) with the first content click of the consent banner studies (second web page interaction, solid blue), we observe high additional delays caused by the required consent banner interactions before being able to interact with the web page. In our study, consent banners cause users to take about 2.9 s on average and 3.1 s on median longer to perform the first content interaction. Note that these delays also depend on the loading times of varying web page elements as well as on user behavior, so our additional delays are not generally valid. Nonetheless, our results demonstrate that consent banners may lead to additional delays which should be kept in mind when optimizing web page loading times and subsequently web QoE.

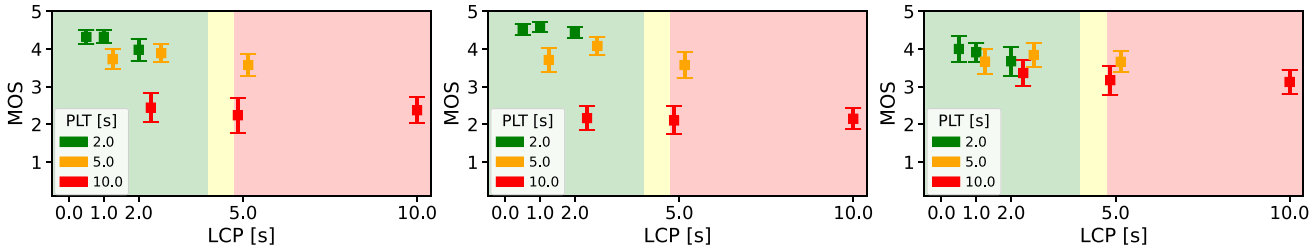
Correlations between ratings, click behavior, and consent banner metrics

Now, we consider the correlations between the user ratings, consent banner related metrics, and click behavior as depicted in Fig. 17.

Regarding the ratings, we can first of all observe that the speed and the experience rating correlate strongly positive (0.78), while the annoyance rating and the experience rating correlate moderately negative (−0.50). Further, we can see that PLT and speed rating correlate strongly (−0.72), while experience ratings correlate moderately to strong with PLT (−0.58), which is lower than in the previous studies. This can again be explained by the fact that the actual PLT was often obfuscated due to suddenly loading consent banners, and thus, users experienced PLT differently than in our previous studies. With respect to the other consent banner related metrics, we observe no strong correlations between them and the ratings. In particular, the ratings and the remaining time after consent banner click (PLT-BC) are also not correlated at all. Due to the relationships of speed rating, PLT, and consent banner loading time, consent banner and time related metrics (BT and BC) are naturally moderately correlated, too.

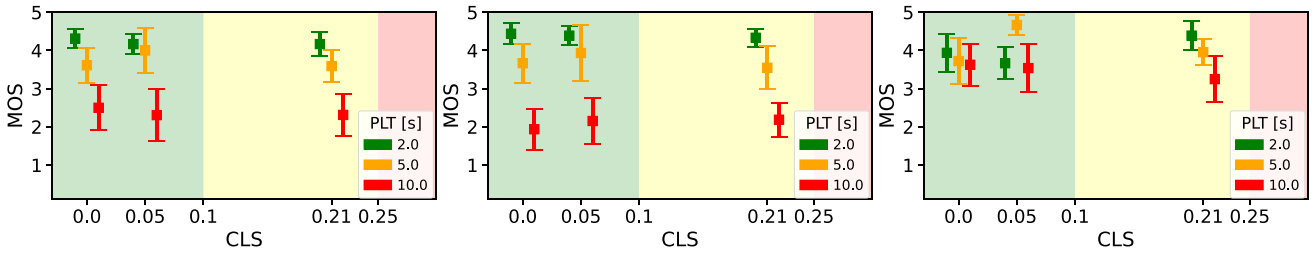
Relation between consent banners and CWV

Finally, we relate the consent banners and the CWV. Here, we consider only LCP and CLS since no artificial input delays had been added to the consent banner buttons during



(a) Overall experience (b) Loading speed (c) Banner annoyance

Fig. 18 MOS for the three QoE dimensions assessed at different LCP and PLT settings



(a) Overall experience (b) Loading speed (c) Banner annoyance

Fig. 19 MOS for the three QoE dimensions assessed at different CLS and PLT settings

these studies. Since each type of consent banner is always the largest element in the viewport, the LCP simply corresponds to the appearance time of the consent banner. CLS is for both *modal* and *interstitial* zero because these consent banners are overlays on top of a page, thus not causing a layout shift. For *header*, there is a layout shift towards the bottom of the page and CLS here depends on the height of the consent banner and the point in time at which the consent banner appears. The appearance of the consent banner (BT) has to be considered in relation to the CWeQS parameters FP, TTText, TTImage, and PLT, since these parameters are decisive for the amount of elements visible in the viewport. Due to equal spacing of these parameters in our study, PLT and BT are sufficient to compute CLS. For BT equal to *PLT*/2, we obtain a CLS of 0.05 and for BT equal to PLT, we obtain a CLS of 0.21.

The relationship between LCP and MOS for the consent banner studies is illustrated in Fig. 18. The figure shows again the MOS for each rating along with the 95% confidence intervals colored according to the PLT. Identical to the CWV page loading studies, we observe that LCP does not affect the MOS for the experience and speed rating at all and that the conditions are stable across Google's recommendations. Both the ratings for experience and speed behave similar and depend only on PLT. For the annoyance rating, we identify neither an influence of LCP nor an influence of PLT. This can be explained by the fact that the question for annoyance has not been tailored towards loading speed, but instead more towards UX.

In the same manner, the effect of CLS on MOS for the consent banner studies is depicted in Fig. 19. Again, we observe that only PLT affects the MOS and that the MOS is stable across Google's recommendations. Interesting though is that for a PLT of 5 s and a CLS of 0.05, participants were least annoyed. This suggests that users prefer consent banners (and here also layout shifts) not too early and not too late if the PLT is sufficiently high, but not too high.

Summarizing, we nonetheless observe that the CWV also do not affect web QoE in terms of MOS when using consent banners.

Discussion

As regards the relationship between the CWV and web QoE, the main finding of our page loading studies is that the CWV do not seem to be good indicators for web QoE in terms of MOS, despite the fact that the CWV are influenced by site loading and rendering behavior. In the different experimental conditions, user QoE ratings depended only on PLT and SI, respectively. In this context, the IQX and WQL hypotheses from previous work [9, 12, 13] also apply to relationship between MOS and SI, which confirms the validity of our measurements. These results

consequently lead us to the question why Google's recommendation for good, moderate, and poor experience for the CWV are not at all reflected in our subjective measurements. After all, Google also relied on human perception and Human-Computer Interaction research to establish these recommendations [41].

A key reason might be differences in overall study designs and data collection approaches. While Google relied on field data focusing on engagement [41], we performed crowdsourcing studies, in which users were not able to stop a web page load without quitting the study and losing their progress. Thus, our studies evaluated instantaneous user opinion ratings, while Google focused more on (longer term) user behavior. These differences of study setups seem to influence the results significantly. As a consequence, we cannot rule out a potential influence of the CWV on the MOS, that we might have not been able to detect due to our study design. Further, we observed a linear relationship between PLT and MOS in our study, which contradicts previous studies. We explain these differences also mainly through the study design. In other studies, participants usually waited passively for the completion of the web page load, while in our studies participants actively searched for the target elements, which they were instructed to click. As a consequence, they experienced the web load in a different context compared to previous studies. Finally, we have tested only three PLTs (2 s, 5,s and 10 s), which strictly limits the general validity of our finding of a linear relationship between PLT and MOS here.

Our follow-up studies focused on the influence of consent banners and investigated the relationship between banner type, web QoE and the CWV, too. Here we again observed that the CWV, this time determined by the rendering process of consent banners, do not correlate with web QoE ratings, which corroborates the findings of the CWV page loading studies. Nonetheless, a comparison of the ratings from the CWV page browsing studies with those from the consent banner studies reveals that web QoE is indeed affected by consent banners. Specifically, consent banners appearing during the page load seem to be beneficial when experiencing higher PLTs as participants became more tolerant to slower page loading. In contrast, consent banners seem to negatively affect web QoE when pages load rapidly. We explain these findings in the following way: As shown above, participants take significantly longer to interact with content if they have to handle a consent banner first. Thus, the delay of accepting or rejecting a consent banner should be considered along with or even on top of the PLT. Rapid page loads with short PLTs may be therefore perceived longer due to suddenly appearing consent banners, which have to be handled first. In contrast to the PLT, the time of appearance of the consent banner did not affect web QoE at all.

Finally, with respect to the type of consent banner, we observed that participants prefer unobtrusive and acentric consent banners like *header*, even though it takes them more time to react to acentric consent banners.

Our consent banner studies were limited in the way that we used very simple consent banners featuring only a small amount of content. Consent banners nowadays are often confusing, sometimes even offering nested configuration options. According to [8], cookie consent banners from third-party providers usually affect web page performance stronger than custom built consent banners, too. In our studies, we did not consider third-party consent banners at all. Further, users were obliged to click on the appearing consent banners before being able to interact with the actual web page. This is in many cases no necessity in reality, but may be an important factor with respect to web QoE. Last but not least, our consent banner studies were limited by the fact that we could test only a few parameter combinations, e.g., only three different PLTs, banner types, and banner appearance times, in our studies. More exhaustive studies are thus required in the future.

Conclusion

In this article, we related Google's Core Web Vitals (CWV) to web QoE by analyzing the results of a series of objective measurement and subjective crowdsourcing campaigns. We presented CWeQS, a novel open-source study framework that we used to perform the subjective web QoE crowdsourcing studies discussed in this work. The framework allows for exerting full control over the loading behavior of custom web pages, a prerequisite for conducting CWV-related experiments. Our objective measurements based on Google Lighthouse revealed that only the LCP metric is actually affected by network conditions, while FID and CLS behave differently for each web page, depending on factors like specific page design and implementation. These findings suggest that accurate in-network monitoring of the CWV on service provider level might actually be very challenging to implement in practice.

Our subjective study results suggest that the CWV metrics do not seem to correlate with web QoE at all. In addition, no QoE impact of Google's recommendations for poor, moderate, and bad CWV scores could be detected. Instead, PLT and SI proved to remain superior indicators for web QoE. This is a surprising result that raises questions regarding the causes behind such discrepancies. We primarily explain these results discrepancies between our studies and Google's CWV-related work with differences in terms of research designs and means of data collection.

Consequently, our next steps for upcoming research include a direct comparison with Google studies by adjusting CWeQS and our CWV studies towards parallel assessment of user engagement and web QoE. In the future, we

also plan to utilize techniques from explainable AI (XAI) to analyze and model the collected data to avoid the introduction of modelling preferences or biases [42].

In addition, we investigated the impact of consent banners on web QoE and the CWV because consent banners represent a content category which has been severely neglected by experience research so far, despite their omnipresence on the web. In dedicated subjective crowdsourcing studies based on CWeQS, we emulated three different types of consent banners, which appeared at different points in time during a page load. Our evaluation results show that consent banners can affect web QoE in both directions in terms of MOS, depending on the PLT. On the one hand, consent banners may lead to slightly more tolerant users when it comes to higher PLTs, while on the other hand, experience slightly decreases with consent banners for fast page loads. In alignment with these findings, we also reveal that the additional time required by users to interact with consent banners should be considered along with the PLT as it may directly influence web QoE. Our results also suggest that unobtrusive and acentric consent banners are perceived the best, even though it takes slightly more time to interact with them. Finally, we reinforced our findings of the CWV page loading studies since we again found no correlation between the CWV, this time determined by consent banners, and web QoE in terms of MOS.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was partly funded by Deutsche Forschungsgemeinschaft (DFG) under grant SE 3163/3-1, project number: 500105691. The authors alone are responsible for the content.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

Ethical approval There are no potential conflicts of interest. This research involved human participants. Informed consent was obtained from participants.

References

1. Baraković S, Skorin-Kapov L (2017) Survey of research on quality of experience modelling for web browsing. Qual User Exp 2:1–31
2. Hossfeld T, Metzger F, Rossi D (2018) Speed index: relating the industrial standard for user perceived web performance to web QoE. In: 2018 Tenth international conference on quality of multimedia experience (QoMEX), pp 1–6
3. da Hora DN, Asrese AS, Christophides V, Teixeira R, Rossi D (2018) Narrowing the gap between QoS metrics and web QoE using above-the-fold metrics. In: 2018 19th passive and active measurement conference (PAM), pp 31–43
4. Jahromi HZ, Delaney DT, Hines A (2020) Beyond first impressions: estimating quality of experience for interactive web applications. IEEE Access 8:47741–47755
5. Chrome Developers, Web Vitals. Accessed 15 Feb 2022. https://web.dev/learn-web-vitals/
6. Pochat VL, Van Goethem T, Tajalizadehkhoob S, Korczyński M, Joosen W (2018) Tranco: a research-oriented top sites ranking hardened against manipulation. arXiv preprint arXiv:1806.01156
7. Wehner N, Amir M, Seufert M, Schatz R, Hoßfeld T (2022) A vital improvement? relating Google's Core Web Vitals to actual web QoE. In: 2022 14th international conference on quality of multimedia experience (QoMEX), pp 1–6
8. Katie Hempenius: Best practices for cookie notices. Accessed 15 Nov 2022. https://web.dev/cookie-notice-best-practices/
9. Ibarrola E, Taboada I, Ortega R (2009) Web QoE evaluation in multi-agent networks: validation of ITU-T G. 1030. In: 2009 Fifth international conference on autonomic and autonomous systems, pp 289–294
10. Egger S, Hossfeld T, Schatz R, Fiedler M (2012) Waiting times in quality of experience for web based services. In: 2012 Fourth international workshop on quality of multimedia experience, pp 86–96
11. International Telecommunication Union (2009) ITU-T recommendation G.1030: estimating end-to-end performance in IP networks for data applications
12. Fiedler M, Hossfeld T, Tran-Gia P (2010) A generic quantitative relationship between quality of experience and quality of service. IEEE Netw 24(2):36–41
13. Egger S, Reichl P, Hoßfeld T, Schatz R (2012) "Time is bandwidth"? Narrowing the gap between subjective time perception and quality of experience. In: 2012 IEEE international conference on communications (ICC), pp 1325–1330
14. Sackl A, Casas P, Schatz R, Janowski L, Irmer R (2015) Quantifying the impact of network bandwidth fluctuations and outages on web QoE. In: 2015 Seventh international workshop on quality of multimedia experience (QoMEX), pp 1–6
15. Asrese AS, Eravuchira SJ, Bajpai V, Sarolahti P, Ott J (2019) Measuring web latency and rendering performance: method, tools, and longitudinal dataset. IEEE Trans Netw Serv Manag 16(2):535–549
16. Rajiullah M, Lutu A, Khatouni AS, Fida M-R, Mellia M, Brunstrom A, Alay O, Alfredsson S, Mancuso V (2019) Web experience in mobile networks: lessons from two million page visits. In: The world wide web conference, pp 1532–1543
17. Varela M, Skorin-Kapov L, Mäki T, Hoßfeld T (2015) QoE in the web: a dance of design and performance. In: 2015 Seventh international workshop on quality of multimedia experience (QoMEX), pp 1–7
18. Varela M, Mäki T, Skorin-Kapov L, Hoßfeld T (2013) Towards an understanding of visual appeal in website design. In: 2013 Fifth international workshop on quality of multimedia experience (QoMEX), pp 70–75

19. Park S, Choi Y, Cha H (2021) WebMythBusters: an in-depth study of mobile web experience. In: INFOCOM, pp 1–10
20. Kelton C, Ryoo J, Balasubramanian A, Das SR (2017) Improving user perceived page load times using gaze. In: NSDI, vol 17, pp 545–559
21. Varvello M, Blackburn J, Naylor D, Papagiannaki K (2016) Eyeorg: a platform for crowdsourcing web quality of experience measurements. In: Proceedings of the 12th international on conference on emerging networking experiments and technologies, pp 399–412
22. Salutari F, Da Hora D, Varvello M, Teixeira R, Christophides V, Rossi D (2020) Implications of the multi-modality of user perceived page load time. In: MedComNet, pp 1–8
23. Guse D, Schuck S, Hohlfeld O, Raake A, Möller S (2015) Subjective quality of web page loading: the impact of delayed and missing elements on quality ratings and task completion time. In: 2015 Seventh international workshop on quality of multimedia experience (QoMEX), pp 1–6
24. Strohmeier D, Jumisko-Pyykkö S, Raake A (2012) Toward task-dependent evaluation of web-QoE: free exploration vs."Who ate what?". In: 2012 IEEE globecom workshops, pp 1309–1313
25. Saverimoutou A, Mathieu B, Vaton S (2019) A 6-month analysis of factors impacting web browsing quality for QoE prediction. Comput Netw 164:106905
26. Yu N, Kong J (2016) User experience with web browsing on small screens: experimental investigations of mobile-page interface design and homepage design for news websites. Inf Sci 330:427–443
27. International Telecommunication Union (2013) ITU-T recommendation G.1031: QoE factors in web browsing
28. International Telecommunication Union (2013) ITU-T recommendation P.1501: subjective testing methodology for web browsing
29. Seufert M, Zach O, Slanina M, Tran-Gia P (2017) Unperturbed video streaming QoE under web page related context factors. In: 2017 Ninth international conference on quality of multimedia experience (QoMEX)
30. Seufert A, Schweifler R, Poignée F, Seufert M, Hoßfeld T (2022) Waiting along the path: how browsing delays impact the QoE of music streaming applications. In: 2022 14th international conference on quality of multimedia experience (QoMEX)
31. Kretschmer M, Pennekamp J, Wehrle K (2021) Cookie banners and privacy policies: measuring the impact of the GDPR on the web. ACM Trans Web (TWEB) 15(4):1–42
32. Rasaii A, Singh S, Gosain D, Gasser O (2023) Exploring the Cookieverse: a multi-perspective analysis of web cookies. In: Passive and active measurement: 24th international conference, PAM 2023, virtual event, 21–23 Mar 2023, Proceedings, pp 623–651
33. Traverso S, Trevisan M, Giannantoni L, Mellia M, Metwalley H (2017) Benchmark and comparison of tracker-blockers: should you trust them? In: 2017 Network traffic measurement and analysis conference (TMA), pp 1–9
34. Muzamil M, Khan A, Hussain S, Jhandir MZ, Kazmi R, Bajwa IS (2021) Analysis of tracker-blockers performance. Pak J Eng Technol 4(1):184–190
35. Jha N, Trevisan M, Vassio L, Mellia M (2022) The internet with privacy policies: measuring the web upon consent. ACM Trans Web (TWEB) 16(3):1–24
36. Habib H, Li M, Young E, Cranor L (2022) "Okay, whatever": an evaluation of cookie consent interfaces. In: Proceedings of the 2022 CHI conference on human factors in computing systems, pp 1–27
37. Philip Walton: Total blocking time (TBT). Accessed 16 Nov 2022. https://web.dev/tbt/
38. De Leeuw JR (2015) jsPsych: a JavaScript library for creating behavioral experiments in a web browser. Behav Res Methods 47:1–12
39. Hoßfeld T, Hirth M, Redi J, Mazza F, Korshunov P, Naderi B, Seufert M, Gardlo B, Egger S, Keimel C (2014) Best Practices and Recommendations for Crowdsourced QoE-Lessons learned from the Qualinet Task Force "Crowdsourcing"
40. International Telecommunication Union (2008) ITU-T recommendation P.910: subjective video quality assessment methods for multimedia applications
41. Sagoo A, Sullivan A, Sekhar V, The science behind web vitals. Accessed 30 Mar 2022. https://blog.chromium.org/2020/05/the-science-behind-web-vitals.html
42. Wehner N, Seufert A, Hoßfeld T, Seufert M (2023) Explainable data-driven QoE modelling with XAI. In: 2023 15th international conference on quality of multimedia experience (QoMEX), pp 1–6