A vital improvement? Relating Google's core web vitals to actual web QoE

Nikolas Wehner, Monisha Amir, Michael Seufert, Raimund Schatz, Tobias Hoßfeld

Angaben zur Veröffentlichung / Publication details:

Wehner, Nikolas, Monisha Amir, Michael Seufert, Raimund Schatz, and Tobias Hoßfeld. 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), 5-7 September 2022, Lippstadt, Germany*, edited by Jan-Niklas Voigt-Antons, Oliver Hohlfeld, Florian Metzger, and Raimund Schatz, 1–6. Piscataway, NJ: IEEE. https://doi.org/10.1109/qomex55416.2022.9900881.

A Vital Improvement? Relating Google's Core Web Vitals to Actual Web QoE

Nikolas Wehner*, Monisha Amir*, Michael Seufert*, Raimund Schatz†, Tobias Hoßfeld*
*University of Würzburg, Institute of Computer Science, Würzburg, Germany, name.surname@informatik.uni-wuerzburg.de
†AIT Austrian Institute of Technology, Vienna, Austria, raimund.schatz@ait.ac.at

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. 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. In these studies, we use CWeQS, a customized framework to emulate the entire web page loading process, and ask users for their experience while controlling the Core Web Vitals. Our results suggest that the Core Web Vitals have less predictive value for web QoE than expected and that page loading times remain the main influence factor in this context.**

I. 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 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 work, we focus on the network-influenced aspects of web QoE, and thus, analyze the relationship between CWV and web QoE by asking: *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.

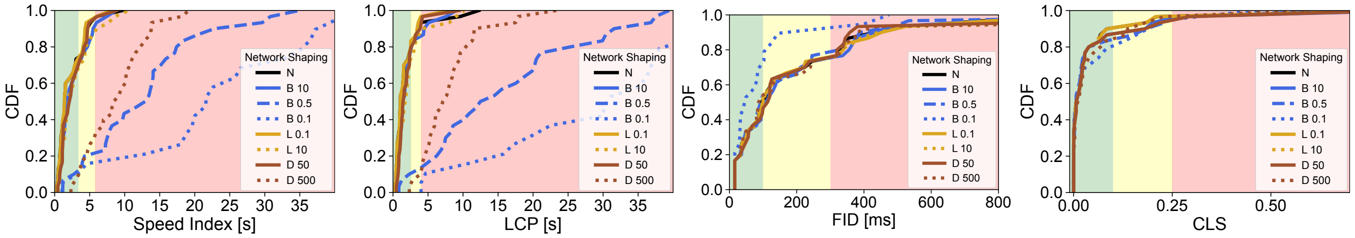
Our objective measurements are performed using Google's Lighthouse and the top 50 Tranco web pages [6]. In particular, we analyze the sensitivity of the CWV in the network by emulating various QoS conditions. These objective measurements are complemented by QoE crowdsourcing studies, in which we emulate 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. We use CWeQS to conduct 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. We contribute by showing that the CWV seem to be less insightful for web QoE than expected and, in particular, inferior to traditional metrics like Page Load Time (PLT) or Speed Index (SI). Our results also indicate that user studies targeting user engagement (as done by Google for the CWV) have to be considered fundamentally different compared to traditional loading time studies.

The remainder of this work is structured as follows: Section II discusses related work. The objective Lighthouse measurements and the corresponding results are presented in Section III. This is followed by the description of our novel study framework CWeQS as well as the description of our performed Core Web Vitals studies in Section IV and Section V. The obtained results from these studies are summarized in Section VI, before Section VII concludes.

II. Related Work

The authors of [7], [8] have shown in early works that loading times, in particular the PLT, are fundamental for estimating web QoE [9]. In the meantime, new metrics focusing on the visible portion of a web page have been proposed, e.g., the Above the Fold Time (ATF) [3] and the Speed Index (SI) [10]. 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



(a) Speed Index.

(b) Largest Contentful Paint.

(c) First Input Delay.

(d) Cumulative Layout Shift.

Fig. 1: CDFs of the 75 percentile of the metrics per web page for the Lighthouse measurement.

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 [11], the WQL hypothesis assumes a logarithmic relationship on a linear ACR scale [7], [12]. Several works have shown network quality fluctuations affect the loading process [13]–[15]. In addition to loading times, web QoE is also influenced by usability [16], aesthetics [17], and device type, i.e., desktop, smartphone, and tablets [1]. The authors of [18] 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 [19] and Eyeorg [20] 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 [19], 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 [21], 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.

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.

III. Objective Measurements with Lighthouse

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 study conditions 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.

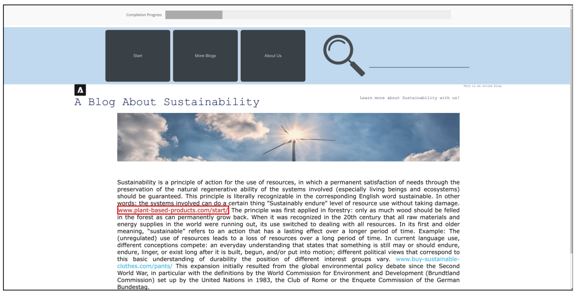
Figure 1 therefore depicts only the results for selected network shaping conditions of the desktop Lighthouse measurements in form of CDFs. As Google recommends that 75% of web page visits should provide a good experience [5], for each web page, we consider the 75 percentile of LCP, FID, CLS, and SI 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 significant different behavior when facing different network conditions (Fig. 1b). In particular, the LCP behaves very similar to the SI (cf. Fig. 1b and Fig. 1a), and both easily end up with poor performance as soon as the network conditions are really bad. This is reasonable as

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

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

[3] https://min.io/



A Blog About Sustainability

Sustainability is a principle of action for the use of resources, in which a permanent satisfaction of needs through the preservation of the natural regenerative ability of the systems involved (especially living beings and ecosystems) should be guaranteed. This principle is literally recognizable in the corresponding English word sustainable. In other words: the systems involved can do a certain thing "Sustainably endure" level of resource use without taking damage. www.plant-based-products.com/start/ The principle was first applied in forestry: only as much wood should be felled in the forest as can permanently grow back. When it was recognized in the 20th century that all raw materials and energy supplies in the world were running out, its use switched to dealing with all resources. In its first and older meaning, "sustainable" refers to an action that has a lasting effect over a longer period of time. Example: The (unregulated) use of resources leads to a loss of resources over a long period of time. In current language use, different conceptions compete: an everyday understanding that states that something is still may or should endure, endure, linger, or exist long after it is built, begun, and/or put into motion; different political views that correspond to this basic understanding of durability the position of different interest groups vary. www.buy-sustainable-clothes.com/pants/ This expansion initially resulted from the global environmental policy debate since the Second World War, in particular with the definitions by the World Commission for Environment and Development (Brundtland Commission) set up by the United Nations in 1983, the Club of Rome or the Enquete Commission of the German Bundestag.

Fig. 2: Example for a custom page (blog page).

both metrics represent loading behavior and indicates that the LCP may act as a proxy for the SI. In contrast, FID and CLS are barely influenced by deteriorating network conditions (cf. Fig. 1c and Fig. 1d). 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.

IV. CWeQS: Crowdsourcing Web QoE Studies

For our web QoE studies we use 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 [22], a JavaScript framework for browser-based studies.

To have 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 with which we 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 pages 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 (as depicted in Figure 2).

To align with best practices for crowdsourced QoE studies [23], 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 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 Figure 2.

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

1) 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 stimuli. After observing all test stimuli, participants are rewarded with a verification code. A training or test stimuli hereby consists of the emulated page load and the subsequent questionnaire, in which participants rate the perceived loading

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

TABLE I: Crowdsourcing study conditions.

CWV Parameter	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

time on the Absolute Category Rating scale [24].

2) 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].

V. Conduct of Core Web Vital Studies

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

1) 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 seconds, we additionally use LCPs of 1 and 1.5 seconds to be able to compare LCP across the different PLTs.

2) 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 seconds.

3) 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 Figure 2, 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 times 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 I, which are again aligned to the recommendations of Google.

VI. Study Results

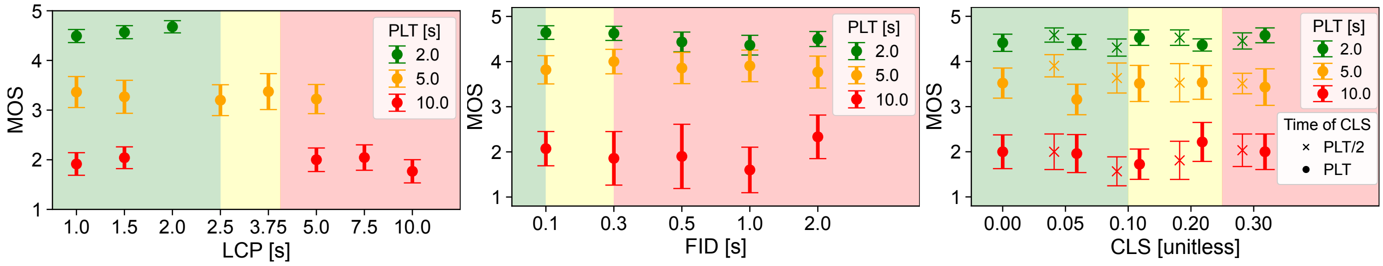
All three studies were conducted in December 2021 and January 2022 using the crowdsourcing platform Microworkers[8]. In pre-studies, we observed that the validation task does not seem to have any influence on the rating behavior of the crowd. Thus, to avoid conflicts with the LCP, we decided to only use hyperlink validation. 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,

[5] https://github.com/WPO-Foundation/visualmetrics

[6] https://www.selenium.dev/

[7] https://www.ffmpeg.org/

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



(a) Largest Contentful Paint.

(b) First Input Delay.

(c) Cumulative Layout Shift.

Fig. 3: Relation of Google's Core Web Vitals (LCP, FID, CLS) to actual web QoE (MOS).

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 ACR scale.

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 3 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 Figure 3c, 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 2s (green), around 3.5 for 5s (yellow), and around 2 for 10s (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.

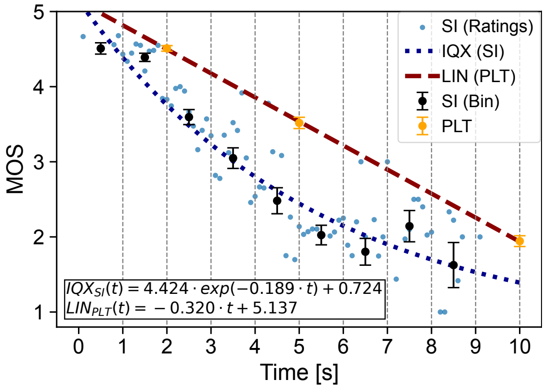


Fig. 4: Relation of PLT and SI to MOS.

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. Both PLT (−0.74) and SI (−0.70) show a high negative Spearman's Rank-Order Correlation Coefficient (SROCC) to the user ratings, which comes as expected considering results of previous QoE studies [2], [8].

Figure 4 visualizes the relationship between averaged user ratings and SI as light blue dots on the continuous SI scale. We also bin the SI in 1s 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 [11] and WQL hypothesis [7], [12] 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 [12], 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 [12].

Discussion: To summarize, we observed in our crowdsourcing studies that, although the CWV are influenced by site loading and rendering behavior, they do not seem to be good indicators for web QoE in terms of MOS. The submitted user ratings depended only on PLT and SI, respectively. While IQX and WQL hypotheses from previous work applied to SI, which confirms the validity of our measurements, we also observed a linear relation between PLT and MOS. 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 measurements. After all, Google also relied on human perception and Human-Computer Interaction research to establish these recommendations [25].

A key reason might be the difference regarding overall study designs and data collection approaches. While Google relied on field data focusing on engagement [25], 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 CVW on the MOS, that we might have not been able to detect due to our study design.

VII. Conclusion

In this work, we related Google's Core Web Vitals (CWV) to web QoE by performing objective Google Lighthouse measurements and subjective crowdsourcing studies. We presented a novel study framework called CWeQS which allows us to measure web QoE while completely controlling the loading behavior of custom web pages and which we used to conduct crowdsourcing studies. Using Google's Lighthouse, we revealed that only LCP is affected by the network, while FID and CLS behave differently for each web page. These findings suggest that accurate in-network monitoring of the CWV could be difficult to implement on service provider level.

In our crowdsourcing studies, we have further shown that the CWV did not correlate well with web QoE. Also, no influence of Google's recommendations for poor, moderate, and bad CWV scores could be observed in the user ratings. Instead, PLT and SI again proved to be the better indicators for web QoE, also confirming the IQX and WQL hypotheses. This is a surprising result provoking questions regarding what causes such discrepancies. We primarily explain these discrepancies between our studies and Google's work on establishing the CWV by the different means of data collection. While Google relied on user engagement measured in the field, we performed crowdsourcing studies targeting both CWV and PLT. These study design differences seem to affect results significantly, a hypothesis to be validated in future work.

As a consequence, our next steps include a direct comparison with Google studies by adjusting CWeQS and our CWV studies such that user engagement is considered.

References

[1] S. Baraković and L. Skorin-Kapov, "Survey of research on quality of experience modelling for web browsing," *Quality and User Experience*, 2017.

[2] T. Hossfeld, F. Metzger, and D. Rossi, "Speed index: Relating the industrial standard for user perceived web performance to web qoe," in *QoMEX*, 2018.

[3] D. N. da Hora, A. S. Asrese, V. Christophides, R. Teixeira, and D. Rossi, "Narrowing the gap between qos metrics and web qoe using above-the-fold metrics," in *PAM*, 2018.

[4] H. Z. Jahromi, D. T. Delaney, and A. Hines, "Beyond first impressions: Estimating quality of experience for interactive web applications," *IEEE Access*, 2020.

[5] "Web Vitals," Accessed: February 15, 2022. [Online]. Available: https://web.dev/learn-web-vitals/

[6] V. L. Pochat, T. Van Goethem, S. Tajalizadehkhoob, M. Korczyński, and W. Joosen, "Tranco: A research-oriented top sites ranking hardened against manipulation," *arXiv preprint arXiv:1806.01156*, 2018.

[7] E. Ibarrola, I. Taboada, and R. Ortega, "Web qoe evaluation in multi-agent networks: Validation of itu-t g. 1030," in *ICAS*, 2009.

[8] S. Egger, T. Hossfeld, R. Schatz, and M. Fiedler, "Waiting times in quality of experience for web based services," in *QoMEX*, 2012.

[9] International Telecommunication Union, "ITU-T Recommendation G.1030 : Estimating End-to-end Performance in IP Networks for Data Applications," 2014.

[10] Google, "Speed index." [Online]. Available: https://sites.google.com/a/webpagetest.org/docs/using-webpagetest/metrics/speed-index

[11] M. Fiedler, T. Hossfeld, and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *IEEE Network*, 2010.

[12] S. Egger, P. Reichl, T. Hoßfeld, and R. Schatz, ""time is bandwidth"? narrowing the gap between subjective time perception and quality of experience," in *ICC*, 2012.

[13] A. Sackl, P. Casas, R. Schatz, L. Janowski, and R. Irmer, "Quantifying the impact of network bandwidth fluctuations and outages on web qoe," in *QoMEX*, 2015.

[14] A. S. Asrese, S. J. Eravuchira, V. Bajpai, P. Sarolahti, and J. Ott, "Measuring web latency and rendering performance: Method, tools, and longitudinal dataset," *TNSM*, 2019.

[15] M. Rajiullah, A. Lutu, A. S. Khatouni, M.-R. Fida, M. Mellia, A. Brunstrom, O. Alay, S. Alfredsson, and V. Mancuso, "Web experience in mobile networks: Lessons from two million page visits," in *WWW*, 2019.

[16] M. Varela, L. Skorin-Kapov, T. Mäki, and T. Hoßfeld, "Qoe in the web: A dance of design and performance," in *QoMEX*.

[17] M. Varela, T. Mäki, L. Skorin-Kapov, and T. Hoßfeld, "Towards an understanding of visual appeal in website design," in *QoMEX*.

[18] S. Park, Y. Choi, and H. Cha, "Webmythbusters: An in-depth study of mobile web experience," in *INFOCOM*, 2021.

[19] C. Kelton, J. Ryoo, A. Balasubramanian, and S. R. Das, "Improving user perceived page load times using gaze," in *NSDI*, 2017.

[20] M. Varvello, J. Blackburn, D. Naylor, and K. Papagiannaki, "Eyeorg: A platform for crowdsourcing web quality of experience measurements," in *CoNEXT*, 2016.

[21] F. Salutari, D. Da Hora, M. Varvello, R. Teixeira, V. Christophides, and D. Rossi, "Implications of the multi-modality of user perceived page load time," in *MedComNet*, 2020.

[22] J. R. De Leeuw, "jspsych: A javascript library for creating behavioral experiments in a web browser," *Behavior research methods*, 2015.

[23] T. Hoßfeld, M. Hirth, J. Redi, F. Mazza, P. Korshunov, B. Naderi, M. Seufert, B. Gardlo, S. Egger, and C. Keimel, "Best practices and recommendations for crowdsourced qoe-lessons learned from the qualinet task force" crowdsourcing"," 2014.

[24] International Telecommunication Union, "ITU-T Recommendation P.910: Subjective Video Quality Assessment Methods for Multimedia Applications," 2008.

[25] "The Science Behind Web Vitals," Accessed: March 30, 2022. [Online]. Available: https://blog.chromium.org/2020/05/the-science-behind-web-vitals.html