WebMythBusters: An In-depth Study of Mobile Web Experience

Seonghoon Park, Yonghun Choi, and Hojung Cha
Department of Computer Science
Yonsei University
Seoul, Republic of Korea
{park.s, y.choi, hjcha}@yonsei.ac.kr

***Abstract*—The quality of experience (QoE) is an important issue for users when accessing the web. Although many metrics have been designed to estimate the QoE in the desktop environment, few studies have confirmed whether the QoE metrics are valid in the mobile environment. In this paper, we ask questions regarding the validity of using desktop-based QoE metrics for the mobile web and find answers. We first classify the existing QoE metrics into several groups according to three criteria and then identify the differences between the mobile and desktop environments. Based on the analysis, we ask three research questions and develop a system, called WebMythBusters, for collecting and analyzing mobile web experiences. Through an extensive analysis of the collected user data, we find that (1) the metrics focusing on fast completion or fast initiation of the page loading process cannot estimate the actual QoE, (2) the conventional scheme of calculating visual progress is not appropriate, and (3) focusing only on the above-the-fold area is not sufficient in the mobile environment. The findings indicate that QoE metrics designed for the desktop environment are not necessarily adequate for the mobile environment, and appropriate metrics should be devised to reflect the mobile web experience.**

Keywords—quality of experience, web page load, web measurements, mobile web, user study

I. INTRODUCTION

Improving the quality of experience (QoE) for users while they access the web is a crucial issue. Users are sensitive to a browser's response, and reluctant to revisit websites where they had an unpleasant experience [1]. To analyze and improve the web QoE, we need to understand what features represent the user experience and then develop techniques for measuring the QoE correctly. QoE is an abstract concept that cannot be quantified directly. Previous studies have suggested various metrics for estimating the improvement or degradation of the QoE by using information collected within a web browser. These metrics quantify the QoE based on several hypotheses derived mostly from the desktop web environment. For example, many metrics focus on the above-the-fold area, while ignoring the below-the-fold area, in the desktop web environment.

Although web usage has rapidly been shifting from desktop to mobile environments [2], web developers often use existing desktop-based QoE metrics to estimate the QoE of the mobile web experience and optimize the performance accordingly [3, 4]. However, there is no guarantee that the existing QoE metrics represent the user experience adequately because the mobile environment is distinctly different from the desktop environment. Without a thorough investigation of the legitimacy of the metrics, the conventional metrics developed in the desktop environment would not be applicable in the mobile environment. Previous studies have shown that the user experience in the mobile web environment has distinctive characteristics. Several studies [5, 6] analyzed the effect of network-related factors on the user experience of mobile web browsing. An interesting finding is that the adoption of network optimization techniques, such as web caching, does not necessarily improve the user experience of the mobile web. Other studies [7, 8] analyzed how the performance of mobile devices affects the user experience of the mobile web. The user experience is known to be affected more by the processing time of web resources on the mobile device than by the network factors. In short, previous work showed differences between the mobile and desktop web environments, but did not reveal how those differences affect the actual user experience during mobile web browsing, and whether existing QoE metrics are still valid.

In this paper, we question the validity of existing QoE metrics for mobile devices and try to find answers with carefully designed and conducted user studies. Several challenges exist for acquiring reliable user experiences in the mobile environment. First, questions should be thoughtfully designed to cover and evaluate various QoE metrics. Because the metrics have different underlying assumptions about the QoE, validating the QoE metrics with only a few questions is a non-trivial task. The questions should be able to validate existing QoE metrics while being orthogonal to each other. Second, the experiment must be conducted to measure the actual user experience in mobile environments, which is an essential requirement for answering the research questions correctly. Existing schemes [9-12] for collecting web experience were mostly developed for desktop environments, and thus, are not readily applicable to mobile environments. Mobile-centric schemes should be developed to show actual web pages on users' devices, control the loading time of each element, and collect reliable responses from users. Considering these issues, we designed three key questions, based on the preliminary experiments on existing QoE metrics applied to desktop and mobile web environments. The questions are answered to validate the applicability of existing QoE metrics to the mobile environment. This requires a user study in real environments with an appropriate tool. To this end, we developed an experimental platform, WebMythBusters, which can easily be distributed to a real user environment, while accurately emulating web page loading to collect various information, such as users' subjective satisfaction and web page loading information.

TABLE I. WEB QOE CLASSIFICATION

Criterion	Group	Metric
Criterion 1. Time point	1. Load completion	PLT, (A)ATF, LVC, TTI, TTC, uPLT
	2. Load start	TTFB, DCL, FP, FCP, FMP, FI
	3. Overall progress	SI, RSI, PSI, BI, OI, RI, MOS
Criterion 2. Measurement target	1. Navigation timing	TTFB, DCL, PLT
	2. Visual change	FP, FCP, FMP, (A)ATF, SI, RSI, BI, OI, PSI, LVC
	3. Interactivity	RI, FI, TTI
	4. Subjective metrics	uPLT, TTC, MOS
Criterion 3. Above-the-fold	1. Above-the-fold	SI, RSI, PSI, BI, OI, (A)ATF, RI, TTC, uPLT
	2. No consideration	TTFB, DCL, PLT, FP, FCP, FMP, FI, TTI, LVC, MOS

The key finding of the study is that the existing QoE metrics developed in the desktop web environment are not effective in the mobile environment. Specifically, the assumption that faster initiation or faster completion of the loading of a web page does not guarantee a better user experience in the mobile environment. In addition, metrics based on the above-the-fold area or conventional methods for calculating visual completeness do not represent the actual QoE in mobile environments. Overall, the findings identify the problems or precautions when evaluating user experience with existing QoE metrics in the mobile web environment. The study results also suggest the need for new QoE metrics that consider the distinctive characteristics of the mobile web environment.

II. WEB QUALITY-OF-EXPERIENCE METRICS

In this section, we describe the existing QoE metrics for the web environment. Table I shows the criteria for classifying the QoE metrics into several groups, and Table II lists the QoE metrics.

A. Web Page Loading Time Point

According to the time point on which a QoE metric focuses, we classify existing QoE metrics into three groups. The QoE metrics in the first group are founded on "how fast the loading of a web page is completed." For example, the commonly used metric of page load time (PLT or onLoad) is the time when the web browser determines that all of the web page's contents have been loaded. In addition to the PLT, several metrics cover the load completion: above-the-fold time (ATF), approximated above-the-fold time (AATF), last visual change (LVC), time to interactivity (TTI), time to click (TTC), and user-perceived page load time (uPLT). The QoE metrics in the second group focus on "how early the loading of a web page starts." For example, Google suggested first paint (FP), first contentful paint (FCP), and first meaningful paint (FMP), and argued that reducing these metrics would improve the user experience. Several metrics define the starting point of the loading process in various ways: time to first byte (TTFB), time to DOMContentLoaded event (DCL), and first CPU idle (FI). The QoE metrics in the third group are related to "how quickly the overall loading process progresses." For instance, the Google-proposed speed index (SI) denotes the average time that is taken to load each visual element within the above-the-fold area, and is calculated using the following equation:

TABLE II. QOE METRICS FOR WEB PAGE LOADS

Name	Description
TTFB	Time to First Byte [13]
DCL	Time to DOMContentLoaded event [13]
PLT	Time to onLoad event; Page Load Time [13]
FP	First Paint [14]
FCP	First Contentful Paint [14]
FMP	First Meaningful Paint [15]
SI	Speed Index [16]
RSI	Real user monitoring (RUM) Speed Index [17]
PSI	Perceptual Speed Index [10]
BI	Byte Index [16]
OI	Object Index [16]
ATF	Above-The-Fold time [18]
AATF	Approximated Above-The-Fold time [18]
LVC	Last Visual Change [11]
RI	Ready Index [19]
FI	First CPU Idle [15]
TTI	Time to Interactive [15]
TTC	Time to Click (used in SpeedPerception) [10]
uPLT	User-perceived Page Load Time [9]
MOS	Mean Opinion Score [20]

$$\text{Index} = \int_0^{t_{end}} \big(1 - f(t)\big)\, dt.$$

The visual completeness (or visual progress) function $f(t) \in [0, 1]$ is the rate at which the content in the above-the-fold area is filled at time t compared to the above-the-fold time t_{end}. The concept is illustrated in Fig. 1. The result of this equation is equal to the area of the shaded region above the curve, which is equal to the average time $t_{average}$ which is taken to load each element. Because measuring the SI requires the costly process of screen recording, various metrics for approximating the SI have been proposed. These include the real user monitoring speed index (RSI), byte index (BI), and object index (OI). The perceptual speed index (PSI), the ready index (RI), and the mean opinion score (MOS) also belong to the third group.

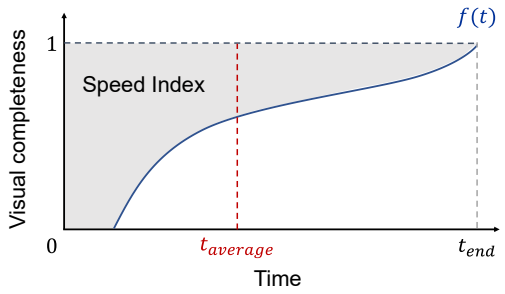


Fig. 1. The visual completeness function of a web page loading and the Speed Index.

B. Measurement Target

The QoE metrics, according to what is measured, fall into four groups: (1) navigation timing, (2) visual change, (3) interactivity, and (4) subjective metrics. The metrics based on the navigation timing are obtained through the Navigation Timing API [13]. The API is an interface for providing developers with performance-related information about web pages, such as the TTFB, PLT, and DCL, measured within web browsers. The visual change, while the web page loads, is an important factor for measuring the user experience. Google submitted the visual feedback-related metric to the W3C standard as the Paint Timing API [14], which is supported in the Google Chrome browser. This API provides several metrics according to the type of visual feedback: FP, FCP, and FMP. Other metrics focusing on visual changes, such as the ATF, AATF, SI, RSI, BI, OI, PSI, and LVC, are obtained by recording videos or approximately estimating visual changes. Allowing users to interact faster with a web page leads to better user experiences because the web pages contain not only visual but also functional elements. If the functional elements are not prepared in time, users cannot use the web page even if it is fully displayed. The QoE metrics considering the loading of the functional elements include the FI, TTI, and RI.

Subjective QoE metrics rest on users' actual responses to the loading processes of web pages. These metrics can be considered the ground truth of the objective QoE metrics discussed above. The subjective QoE metrics include uPLT, TTC, and MOS. The uPLT is the PLT that users actually experience, which can be the ground truth of the PLT. Existing studies [9, 11] measured the uPLT first by recording the above-the-fold area while the web page loaded and then surveyed the time when the user thought the web page had finished loading. The TTC is the time when a user makes a judgment regarding whether a page load is completed. The TTC and the uPLT attempt to acquire the ground truth of the page load time. The MOS measures the QoE by allowing users to score their satisfaction on a 1- to 5-point scale, after the web page loads.

C. Above-the-fold Area

Because only the above-the-fold areas are visible to users during the initial phases of web page loading, visual changes in the below-the-fold areas likely may not affect the user experience. Based on this idea, the above-the-fold time (ATF), the time when the visual change in the above-the-fold area ends, was suggested. Other similar metrics have also been devised based on this idea: the SI, RSI, PSI, BI, OI, AATF, RI, TTC, and uPLT. The development of many relevant metrics indicates that the above-the-fold area has been considered important in the desktop environment.

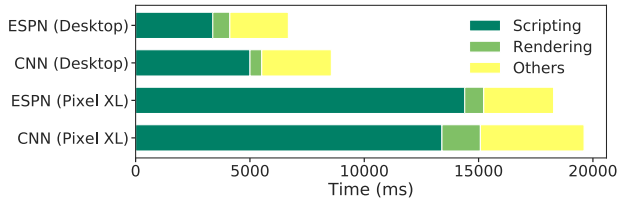


Fig. 2. Computing requirement of web loading processes

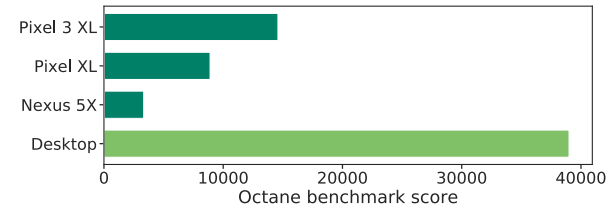


Fig. 3. The result of Octane benchmark

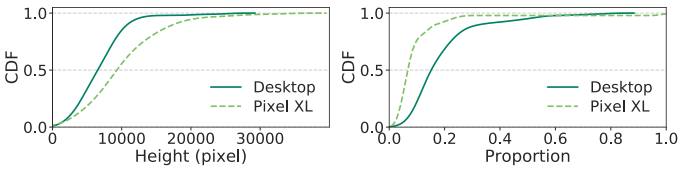
III. Preliminary Experiments

Existing QoE metrics developed for the desktop web environment may not be applicable to a mobile environment. This being likely, we first analyze the difference between the mobile and desktop environments in terms of the performance and the user interface.

A. Performance

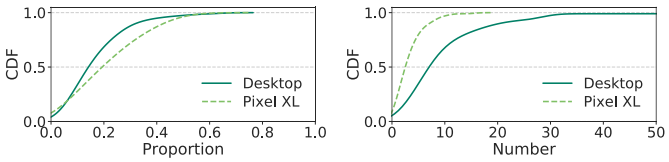
Mobile devices are generally less powerful in computation than desktop devices. The performance bottleneck of web page load in the mobile environment is known to be computation, whereas networking is the case in the desktop environment [4, 6]. To further investigate this issue, we evaluated web page loading on the Pixel XL smartphone and a desktop PC. The Pixel XL device has 4GB RAM with Snapdragon 821 MSM8996 Pro SoC AP (Kyro MP2 2.2 GHz + MP2 1.6 GHz). The desktop PC (hereafter, called a desktop) has 16GB RAM and Intel Core i5-7600 CPU (3.50 GHz). The difference in computing power between the two devices is substantial. The heat and energy issues degrade the performance of the mobile device even further. Both the mobile device and the desktop run Chrome version 76.

We analyzed the loading process of two popular websites, CNN and ESPN, in terms of the computation requirements. Fig. 2 shows the time taken for the web page loading process, mainly, the scripting and the rendering processes. The analysis is conducted with the Chrome DevTools Performance panel [29]. The mobile device is significantly slower than the desktop. Specifically, processing the script takes a larger portion of the page load time in the mobile device. This is due to the difference in computing power between the mobile device and the desktop. We conducted another test to compare the performance of JavaScript processing, in particular. Fig. 3 shows the results of the Octane benchmark [30], running on the desktop and the smartphones (Pixel XL, Pixel 3 XL, and Nexus 5X). The desktop score is approximately three to ten times higher than that of the smartphones. Such a large difference can lead to a non-



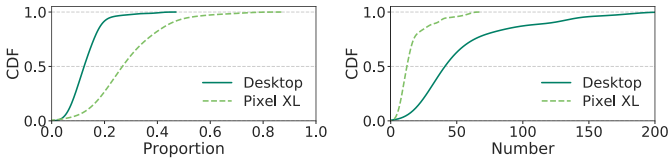
(a) Height of web pages (b) Proportion of above-the-fold areas

Fig. 4. Web page characteristics.



(a) Proportion in area (b) Elements per viewport

Fig. 5. Image areas on desktop and mobile web pages.



(a) Proportion in area (b) Elements per viewport

Fig. 6. Text areas on desktop and mobile web pages.

trivial difference in user experience between the mobile and desktop environments.

B. *User Interface*

One distinctive difference between a mobile device and a desktop is the screen resolution. The resolution of the viewport in desktop web browsers is usually 1920 × 1080, whereas a mobile device such as the Pixel XL has a resolution of 412 × 604. Naturally, the web pages for mobile devices come in narrower and longer layouts, leading to a smaller proportion of the above-the-fold area than that of a desktop. To further analyze the difference concerning the user interface, we investigated Alexa top 50 news sites and 50 sports sites.

Fig. 4(a) shows the cumulative probability distribution (CDF) of the vertical length of selected web pages on the desktop and the Pixel XL. The vertical length of the web pages loaded in the Pixel XL is longer than those loaded on the desktop. Fig. 4(b) shows the CDF of the proportion of the above-the-fold area. The proportion in the Pixel XL is less than half that on the desktop. The difference between the Pixel XL and the desktop indicates that mobile devices have narrower but longer web pages than the desktop.

Similarly, mobile web pages have a larger proportion of images on a page than desktop web pages. Mobile web pages contain fewer images per display than desktop web pages, however. We obtained the image areas in web pages by traversing the Document Object Model (DOM) trees to find image-tagged elements or elements with background images. Fig. 5(a) shows the CDF for the percentage of image areas over the entire web page. The proportion of images on mobile devices is higher than that on desktops. Fig. 5(b) shows the CDF of the number of images in a viewport. Compared to desktop web pages, mobile web pages have fewer images per viewport.

In terms of the text area, the difference is more significant. We obtained text areas by searching elements with innerText tags in leaf nodes from DOM trees. Fig. 6(a) shows the CDF of the size proportion of the text area on the web pages. The proportion of the text areas on the Pixel XL is about twice as large as that on the desktop. This is more significant than the proportion of the image sizes. Fig. 6(b) shows the CDF of the number of text areas contained in a viewport. The Pixel XL viewport contains significantly fewer text areas than the desktop case.

IV. Research Questions

Existing QoE metrics for web page loading are mostly designed in the desktop environment. As previously discussed, the performance and the user interface of the web in mobile environments are distinguished from those in desktop environments. The distinction can lead to different user experiences in web access. To examine the user experience on mobile devices, we ask three research questions regarding web page loading. The research questions are based on the groups in Table I. The three research questions we investigate are summarized in Table III. We focus on the visual change group in Criterion 2, and specifically create questions in conjunction with the groups in Criterion 1 (RQ1, RQ2) and Criterion 3 (RQ3). The questions are about how closely each group is related to the user's real experience.

We concentrate on the visual change group in Criterion 2. The metrics in the navigation timing group have been proved to exhibit a low correlation with the user's real experience [11, 19]. The metrics in the interactivity group are meaningful, but they are too complicated to measure, and few works have adopted these metrics. Meanwhile, the visual changes in web page loading are considered a crucial aspect of the user experience. We use subjective metrics, such as the uPLT and the MOS, as the ground truth for the experiments.

RQ1. Are the metrics that indicate load completion or load start sufficient for estimating actual QoE? In the desktop environment, faster completion or faster initiation of the loading of the web page is assumed to provide a higher quality of experience. However, little work has validated this assumption in the mobile environment. If the assumption holds in the mobile environment, the metrics in the load completion group and the load start group in Criterion 1 are still appropriate for measuring the quality of the experience. Otherwise, the existing works that aim to reduce a page load time or make the start of a page load earlier have few implications in the mobile web environment.

RQ2. Should the visual progress be acquired in the same way as for the desktop environment? The visual progress should be accurately measured when discussing the metrics belonging to the third group in Criterion 1. A typical method is to use pixel comparison, a technique often employed in the SI. Because web pages in the mobile environment tend to have simple structures, fewer elements exist within a viewport. In other words, compared to the web pages in the desktop environment, the weight of each element on a web page is larger in the mobile environment. That is, a text element tends to occupy a large portion of the web pages in the mobile environment. The metric designed for the desktop environment may underestimate the importance of the text elements in the

TABLE III. THE RESEARCH QUESTIONS

No.	Group	Desktop	Mobile
RQ1	Load completion, Load start (Criterion 1)	Quick completion or quick start of a web page load leads to good QoE [9, 12, 15].	?
RQ2	Overall progress (Criterion 1)	The visual completeness can be calculated by comparing pixels to the last frame [16].	?
RQ3	Above-the-fold (Criterion 3)	Focusing on the above-the-fold area is adequate for estimating QoE [18].	?

mobile environment. If the weight is different in the mobile environment, then the method for acquiring the visual progress must be redefined accordingly.

RQ3. Can the below-the-fold area be ignored? As discussed in Section II.C, many metrics focus on the above-the-fold area in desktop environments. The metric originates from the idea that the portion of the page that the user is currently viewing must be loaded first. Mobile web pages are often narrower and vertically longer than desktop web pages, resulting in a smaller portion of the above-the-fold area. We should then ask whether the below-the-fold area is meaningful on mobile web pages. If the answer to RQ3 is negative, then any technique attempting to improve the user experience based on only the above-the-fold metrics may not be effective in mobile environments.

V. DESIGN AND IMPLEMENTATION

To answer the research questions, we designed and implemented a platform, called WebMythBusters, which collects and analyzes the user experience in mobile web environments. Several issues exist in developing the platform. Specifically, we should consider (1) how the web page loading process is presented to users, (2) which types of web pages are presented to users, (3) how we collect information about the user experience, and (4) how we obtain reliable responses from users.

Fig. 7 shows the overall structure of WebMythBusters. It consists of two key parts: the Client and the Server. The Client is an application that users employ directly on their devices. The program replays the page loading processes and displays the surveys on the screen. We developed the Client as an extension program for Chrome. Because the mobile version of Chrome does not support extensions, we used the Kiwi browser [21], which is a modified version of Chrome that allows extensions. The Server governs the overall experiment, sending the experiment parameters to the Client, receiving data from the Client, and storing the data.

A. How to show the web page loading process?

The web page loading process refers to the procedure that begins when a user accesses a web page until the web page is ready for use. Some considerations are required to present the web page loading process to users. First, users should experience the web pages actually loaded through web browsers, instead of recorded videos. Especially for RQ3, we need to understand how users perceive the below-the-fold area while the web page loads. Users should be able to observe the below-the-fold area and interact with the web pages. Note that with a recorded video, which was commonly used in previous research, users observe only the above-the-fold area and are unable to scroll through the video. Second, users should experience a visually consistent loading process. For instance, to find an answer to RQ1, the time to the last visual change or the first paint events should occur at

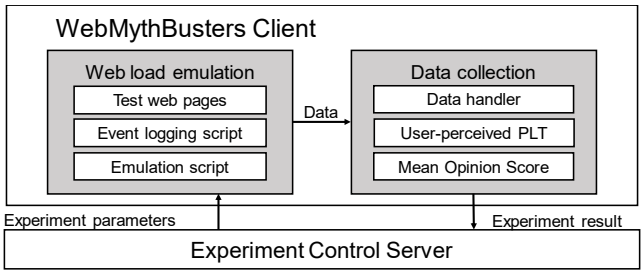


Fig. 7. The WebMythBusters platform.

precisely the same time for the same case, whatever the environment. Unlike recorded videos, actual web page loads allow users to experience different visual changes depending on the environments. We need to minimize the difference that may occur in the web page loading process so that every user experiences a uniform web page loading process.

Considering these requirements, we developed the Client for WebMythBusters. The Client is an emulator that loads web pages on real-world browsers and presents consistent web load progress in various environments. WebMythBusters uses compressed HTML files obtained through SingleFile [22]. The simple structure of HTML files enables the emulation of the web page loading process with little overhead. We wrote scripts to access the DOM and CSS in the HTML files so that certain elements appear on the screen at a particular time point. Thus, users experience scrollable web page loads and consistent visual changes in various environments.

B. Which web pages to present?

When showing the web page loading process to users, we need to decide which web page and what visual changes to present. In addition, the length of the experiment is important. Collecting a large amount of experimental data is helpful, but the longer the experiment, the more likely users are to participate in the experiment unreliably. Considering these issues, we set up experiment parameters, which consist of the URLs of the pages and timing information regarding visual changes. The number of web page loading processes shown to the participants was 60, which consists of 5 websites and 12 different cases per site. We selected five popular websites: CNN, ESPN, Wikipedia, Amazon, and YouTube. ESPN is the only case that contains the advertisement contents in the above-the-fold area.

Table IV shows the 12 cases representing different visual change processes. Each case consists of the values for when a particular area appears on the screen. The FP value indicates when the first paint occurs. The Text value, Image value, Ad value, and BTF value indicate when the text area, image area, advertisement area, and below-the-fold area appear on the screen, respectively. The 12 cases are specifically designed to answer the research questions. For RQ1, we need cases undergoing the same visual changes, but the time to first paint is

TABLE IV. QOE METRICS FOR WEB PAGE LOADS

Case	A	B	C	D	E	F	G	H	I	J	K	L
FP	0	0	0	1	2	2	2	3	4	2	3	3
Text	4	4	4	4	4	4	4	4	4	6	6	9
Image	6	6	6	6	6	6	6	6	6	4	9	6
Ad	8	8	8	8	8	8	8	8	8	8	12	12
BTF	0	5	10	0	0	5	10	0	0	0	0	0

different (cases A, D, E, H, and I). To answer RQ1 and RQ2, we need cases where the time to the last visual change is the same, but the cases undergo different visual changes (cases E, J, K, and L). To answer RQ3, we need cases where the below-the-fold areas appear on the screen at different times (cases A, B, C, E, F, and G).

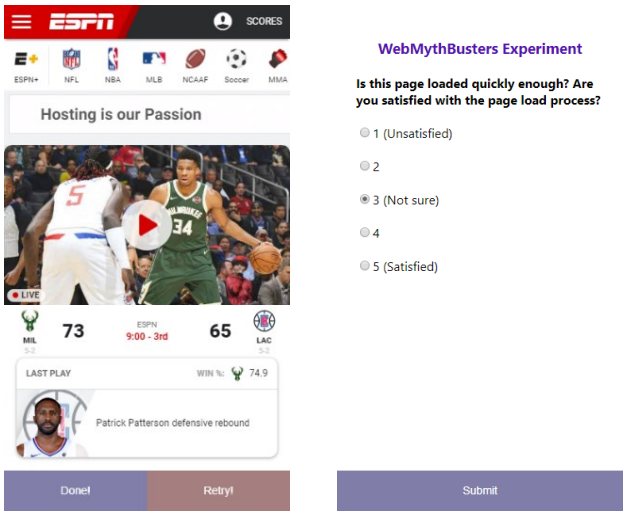
C. How to collect data on the user experience?

WebMythBusters should record data related to the experience of the users, i.e., the participants in the experiment. The participants are asked to answer a survey regarding the uPLT and the MOS. WebMythBusters collects data on user-generated scroll events and collects device information such as screen resolutions. Note that WebMythBusters does not collect any sensitive data, such as web history and personal information.

We define the uPLT slightly differently from previous studies. In previous research, the uPLT was obtained by focusing only on the above-the-fold area and by strictly comparing the frames of the recorded videos. In the present study, the emulated page load process, instead of recorded videos, is presented to participants, and user interactions, such as scrolling, are obtained directly from the participants. In the experiment, the uPLT provides information about when the users perceive the loading of the web page is completed, rather than information about specific time points. Fig. 8(a) shows a screenshot of WebMythBusters when the participants see the web page loading process and answer the survey about the uPLT. The participants click the 'Done' button when the web page load is considered to be complete. After the participants respond to the survey on the uPLT, WebMythBusters displays the MOS-related questionnaire, shown in Fig. 8(b), asking for a 5-point scale answer. At the end of the experiments, we solicit comments from the participants about the overall experimental process.

D. How to obtain reliable responses?

Several methods are used to acquire reliable and faithful responses from participants. We defined hard rules that the participants must follow. For example, we prevented the participants from navigating to other tabs in the browsers, so that the participants would not be distracted from observing the page loads during the experiment. We also defined soft rules that were not explicitly given to the participants. For example, answering the survey about the uPLT before the first paint event is not a reasonable response. This case should certainly be filtered out. Following the concept of the wisdom of the crowd, we considered the majority response of the participants as a pseudo-ground truth. We filtered out the outliers, which are the



(a) uPLT (b) MOS

Fig. 8. The sample questionnaires.

responses that diverged significantly from the pseudo-ground truth, with Tukey's fences [23].

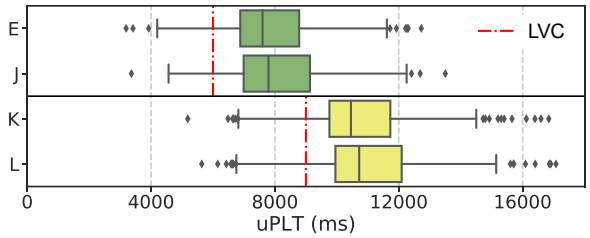
Furthermore, we showed the web page loading process to the participants in a different order. The participants' satisfaction with a web page loading process can vary depending on which process they experienced just before. For example, if a participant experienced a satisfying process seconds before, the current process could be relatively unsatisfactory, and vice versa. WebMythBusters shuffled the sequence of 60 web page loading processes and provided them to the participants.

VI. EVALUATION

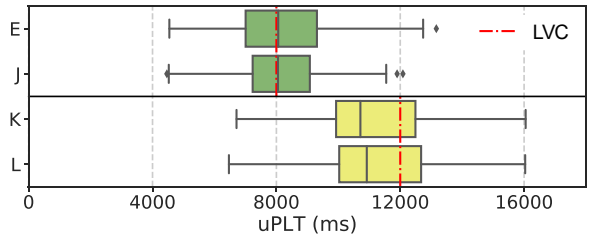
We conducted a user study using WebMythBusters. We recruited 100 participants for the study, mostly college students in their 20s, and collected 6,000 data in total. This user study was approved by the Institutional Review Board (IRB) of our institution.

A. RQ1: Fast Completion and Fast Initiation

We first examined the relationship between the LVC and the uPLT. The LVC is an objective metric for the completion of the loading of a web page, whereas the uPLT is a subjective metric. Fig. 9 shows the effect of the overall speed of the loading of the web page on the uPLT. The analysis was conducted for cases with or without advertising content (i.e., ads) on the pages. For the cases without ads (Fig. 9(a)), the last visual change events occur when the text and image areas are loaded. For cases E and J, the visual changes end at 6 s, which means that the LVC is 6 s. For cases K and L, the visual changes end at 9 s. For cases E and K, the text areas load first, followed by the image areas. For cases J and L, the image areas load first and then the text areas. As shown in the figure, almost all participants feel the web page loading is finished after the last visual changes, meaning that the text and image areas are critical for the uPLT metric. For the cases with ads on the pages (Fig. 9(b)), cases E and J, the visual changes end at 8 s. For cases K and L, the visual changes end at 12 s. In this case, the last visual change events occur when the ad areas are loaded. For cases E and J, half of the participants perceive the web page loading is finished before the last visual

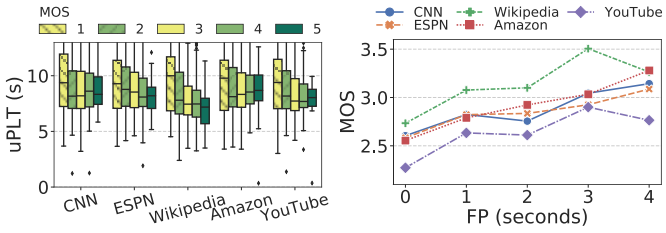


(a) Without advertisement areas



(b) With advertisement areas

Fig. 9. The uPLT distributions in cases E, J, K, and L (without/with ads).



(a) MOS vs. uPLT (b) MOS vs. FP

Fig. 10. Relationships between the MOS with the uPLT and the FP.

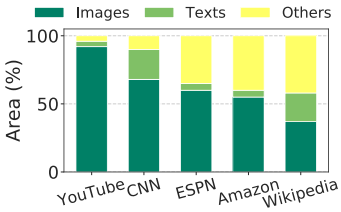


Fig. 11. Text areas vs. image areas.

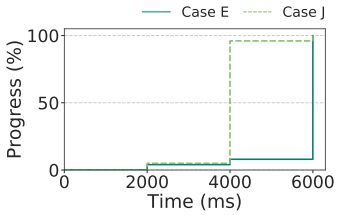


Fig. 12. The visual progress on YouTube (cases E and J).

changes. For cases K and L, more than half of participants perceive the web page loading is finished before the last visual changes. This means that the ad areas of the web pages have little influence on the users' determination of whether the web page loading is finished. In summary, users perceive that the web page is loaded when the text and image areas are loaded, but the ad areas do not have a significant impact. Few QoE metrics in the load completion group have reflected this complicated fact concerning the completion of web page loading.

We also investigated the relationship between the uPLT and the MOS. The uPLT is the ground truth for the user experience of fast completion, and the MOS is the ground truth for the user experience for the overall web experience. By reviewing cases E, J, K, and L, we examine whether a short uPLT leads to a high MOS. Fig. 10(a) indicates the higher the MOS, the smaller the average uPLT in some cases. However, the same uPLT can represent a different MOS, indicating the relationship is not one-to-one. This is why the uPLT is not sufficient for estimating the actual QoE. Fast completion does not always lead to high user satisfaction. Fig. 10(a) also shows that the ad areas do not affect the uPLT. Although ESPN has ads in the above-the-fold area, and the area loads slower than other areas, the corresponding MOS is not degraded. This result implies that the impact of the web page components on the QoE should be considered to correctly measure the user experience.

We also examined the effect of a quick start for the loading of a web page on the QoE by analyzing cases A, D, E, H, and I. Note that the case in which the time to first paint is short has a long interval from the first paint to the other paint. Fig. 10(b) shows the average of the mean opinion scores when the time to first paint is 0, 1, 2, 3, and 4 s. Interestingly, the longer the time to first paint on any web page, the higher the mean opinion score. The user survey comments related to RQ1 help understand the results: "It was more satisfying when the web page was loaded at once." (Several participants submitted similar feedback.) This comment explains why the participants' opinion score was higher, although the start of the loading was slow in Fig. 10(b). The length of time from the first visual change to the second is more important to the user experience than the start time of the loading of the web page. The result means that attention should be paid to estimating the user experience through the time to first paint. Additionally, techniques that reduce only the time to first paint do not improve the mobile web experience.

The findings are as follows:

- Quick completion of a web page load does not always lead to a better QoE.
- A faster start of a web page load can lead to a lower quality of experience.

B. RQ2: Evaluating Visual Progress on the Mobile Web

We examine the proportion of each area in the above-the-fold area on mobile web pages. Fig. 11 shows the percentage of the number of pixels, in the image and text areas, in the above-the-fold area on the five web pages used in the experiment. More pixels are observed in the image area than in the text area. Because the visual progress function in the speed index is calculated by comparing the pixels, quickly loading the image area leads to a good score. Fig. 12 shows two visual progress cases on YouTube (cases E and J). The case where the image area is loaded first (case J) seems faster than the text case (case E). The speed index of case J is 4000 ms, which is significantly better than that of case E, 5760 ms.

To answer RQ2, we compare the cases where the text area appears after the image area (cases J and L), and the reverse cases (cases E and K). Fig. 13 shows the participants' satisfaction. We classify 1 and 2 points of the MOS as negative, 3 points as neutral, and 4 and 5 points as positive. In Fig. 13(a), when the load process for YouTube is observed, the quality of the user experience is low when the SI score is bad (case E). In Fig. 13(d), we obtain the same result when we compare cases K and L. Case K has the same order of showing areas as case E, and case L has the same order of showing areas as case J. The only difference between cases K and E is the loading time of each area, which is the difference between cases L and J.

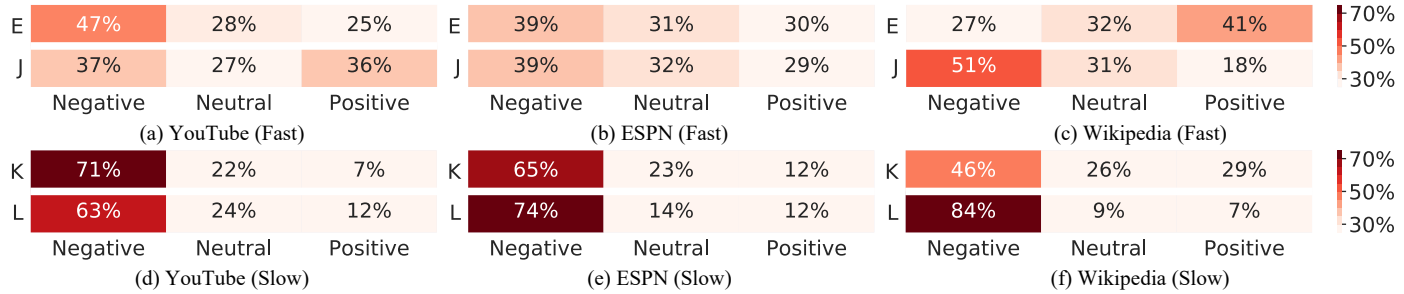


Fig. 13. Relationship between the MOS and the order of showing areas (cases E, J, K, and L).

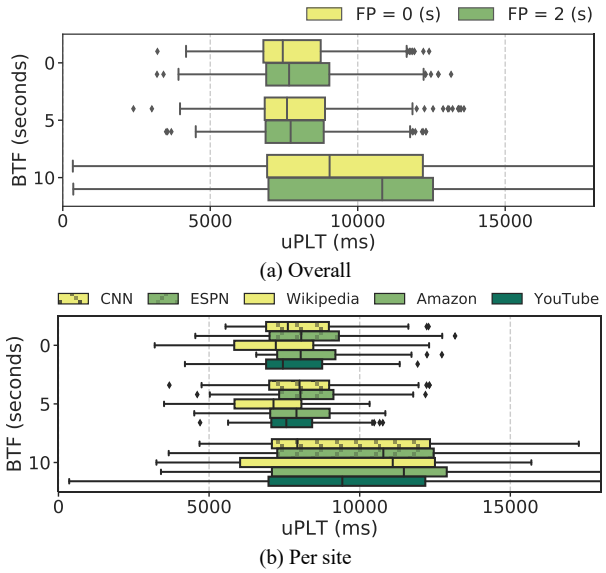


Fig. 14. Distributions of the uPLT when the below-the-fold area appears at 0, 5, and 10 s.

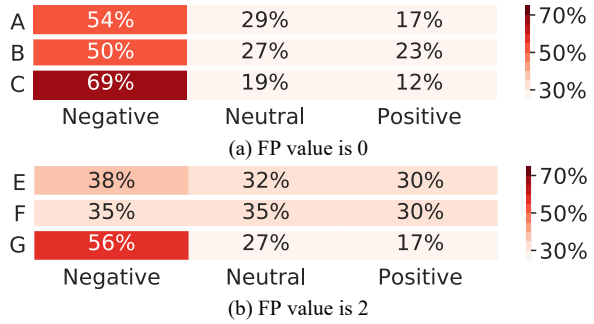


Fig. 15. Distributions of the uPLT when the below-the-fold area appears at 0, 5, and 10 s.

However, in Fig. 13(b), 13(c), 13(e), and 13(f), the participants are more negative about cases J and L than about cases E and K, respectively. Note that ESPN and Wikipedia have larger portions of text areas than YouTube, yet the portions of the text areas are still smaller than those of the image areas. We consider that even if the number of pixels in the text areas is smaller than that in the image areas, the importance of the text areas for user satisfaction is greater than the weight of the image areas. This indicates that it is not appropriate to use the existing methods for calculating the visual progress while web pages load in mobile environments.

The participants' comments related to RQ2 were mostly about what element should be loaded first to make the user more satisfied. The following feedback examples state that loading the text area first is observed to be positive: "I was particularly dissatisfied with the blog text loaded late." "For the documents where the text is important, such as Wikipedia, it was satisfying when the text was displayed first." This feedback explains the results in Fig. 13(b), 13(c), 13(e), and 13(f). However, some feedback was positive about the images loading first: "While loading the YouTube page, it was better when the thumbnail image was displayed first." "In the case of a web page whose main content is a video, the picture being loaded first was satisfying." The comments mostly refer to the YouTube page, which has more images than other websites. This feedback explains the results in Fig. 13(a) and 13(b), and suggests that the weight of each element should be considered differently when measuring user experience.

The findings are as follows:

- The conventional methods for calculating visual progress are not appropriate on the mobile web.
- Text areas are more important than image areas in terms of the actual QoE.
- The speed index underestimates text areas and overestimates image areas.

C. *RQ3: Implication of the Below-the-fold Area*

Thus far, we have looked at cases where the parameters for the below-the-fold area are not diverse. To answer RQ3, we compare cases where the time for loading below-the-fold areas varies (cases A, B, C, E, F, and G). Fig. 14 shows the distribution of the uPLT for the cases where the below-the-fold areas appear at 0, 5, and 10 s. In these cases, the FP is 0 or 2 s, and the ATF is 6 (CNN, Wikipedia, Amazon, and YouTube) or 8 s (ESPN). The figure shows that little change is made to the uPLT when the below-the-fold area appears before the last visual change of the above-the-fold area. However, the uPLT increases significantly when the below-the-fold area appears after the last visual change in the above-the-fold area. Fig. 15 shows that participants respond negatively when the below-the-fold area appears late. This means that users are affected by below-the-fold areas when perceiving the load time.

We also analyze how far participants scroll down while observing the web page loading. Fig. 16 shows the distribution of the scrolled length as a ratio of the device's height. Note that we do not explicitly request the participants to scroll, but the participants themselves scroll while observing the web page

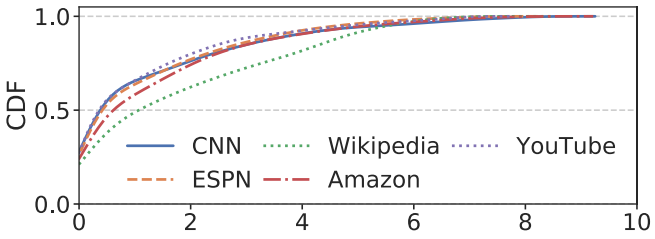


Fig. 16. The distribution of how far the participants scroll down with respect to the viewport of their devices.

loading process. About 80% of the data have scroll events, and the median of the data is 0.5. This result indicates that the below-the-fold area should be considered in terms of the QoE.

The findings are as follows:

- The below-the-fold area should not be ignored in terms of the QoE.
- Users are not satisfied when the below-the-fold area loads slowly.
- Users scroll while waiting for a web page to load completely.

VII. Related Work

Various works have attempted to understand how web page loading in the mobile environment is different from that in the desktop environment. Nikravesh et al. [24] found the cases where inefficient content delivery network (CDN) occurs and the fact that the processing power caused a bottleneck, by analyzing the loading processes of mobile web pages. Nejati and Balasubramanian [6] proposed WProf-M to analyze web page loading on mobile devices. WProf-M showed that computation is the bottleneck while a web page is loading on a mobile device, whereas the bottleneck on a desktop is typically caused by networking. Vesuna et al. [5] showed that caching in a mobile browser is not as effective as in a desktop browser because caching remedies only the delays caused by networking. Dasari et al. [7] analyzed how the performance of the device affects the Internet QoE. Jun et al. [25] analyzed how Google's AMP technology affected web page loads and showed that reducing the complexity of web pages greatly enhanced the load time. Rajiullah et al. [26] analyzed mobile network experiences using various user experience metrics. Although many previous works have shown what characteristics affect the user experience in the mobile web environment, the researchers employed metrics designed for the desktop environment.

Meanwhile, research has been conducted to analyze the QoE of web page loading. Egger et al. [27] asked participants to answer how long their devices took to load a web page. The perceived completion of the loading of a web page was substantially different from the PLT recorded by the browser. Varvello et al. [9] designed Eyeorg, a crowdsourcing platform that surveys the subjective experience of the web page load time. Eyeorg focused on how to obtain a reliable response from users. Gao et al. [10] and Wang et al. [12] designed crowdsourcing platforms named SpeedPerception and Kaleidoscope, respectively, to evaluate the user experience of the desktop web environment. SpeedPerception improved the SI, while Kaleidoscope evaluated web features. Kelton et al. [11] showed that the SI or the PLT has a low correlation with the uPLT. Other works attempted to measure the QoE. Bocchi et al. [20] used a 5-point scale MOS, and Salutari et al. [28] used a 3-point scale MOS to measure the QoE level. Although these works tried to measure the actual QoE based on user studies, the studies did not cover the experience in the mobile environment.

VIII. Discussion

This work contributes to the measurement of the quality of users' experience on mobile websites in two aspects. First, we showed that the performance and the UI structure of mobile web pages are different from those of desktop web pages, and these differences should be reflected in the correct measurement of the QoE on mobile web pages. The correlation between the user experience and the characteristics of mobile websites has not been sufficiently investigated in existing studies. We uncovered this correlation through experiments in real user environments, and presented factors that should be considered when designing an appropriate QoE metric for mobile web pages. Second, this experimental tool is not only easy to distribute to experimental participants but can also perform experiments that were not possible with existing video-recording methods. Recent user studies often use crowdsourcing platforms, and this tool is easily distributed to participants in experiments through these platforms. In addition, the tool can be extended to record the actual web page loading processes on users' devices and replay them on other users' devices. This way, various studies that had been conducted only in a controlled environment can be conducted on the mobile devices of real users.

IX. Conclusion

In this work, we questioned the validity of existing web-related QoE metrics for mobile devices. The analysis showed that many existing metrics that are designed for the desktop environment are not necessarily appropriate for mobile devices. This finding has several new implications. Optimizing mobile web pages using existing metrics may result in a sub-optimal effect on the user experience. We certainly need new metrics that can measure the user experience accurately in the mobile web environment. In particular, this research suggests two findings that should be considered to design a QoE metric that reflects user satisfaction better: priority among the components of the web page and the importance of the below-the-fold area, which has been undervalued. Building on the new metrics, the mobile web will be optimized differently from the desktop web. We hope that this research provides a step in the right direction for optimizing the mobile web environment.

Acknowledgments

This work was supported by Next-Generation Information Computing Development Program funded by the Ministry of Science and ICT (Grant No. NRF-2017M3C4A7083677), National Research Foundation of Korea(NRF) (Grant No. NRF-2019R1A2C2004619), and Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT)(No. 2018-0-00532, Development of High-Assurance (≥EAL6) Secure Microkernel).

REFERENCES

[1] Find out how you stack up to new industry benchmarks for mobile page speed, https://www.thinkwithgoogle.com/marketing-resources/data-measurement/mobile-page-speed-new-industry-benchmarks/.

[2] Mobile web usage overtakes desktop for first time, https://tinyurl.com/zsuasva.

[3] R. Netravali, and J. Mickens, "Prophecy: Accelerating mobile page loads using final-state write logs," 15th {USENIX} Symp. Networked Systems Design and Implementation ({NSDI} 18), USENIX Association, 2018, pp. 249-266.

[4] V. Ruamviboonsuk, R. Netravali, M. Uluyol, and H. V. Madhyastha, "Vroom: Accelerating the mobile web with server-aided dependency resolution," Proc. Conf. ACM Special Interest Group on Data Communication, ACM, 2017, pp. 390-403.

[5] J. Vesuna, C. Scott, M. Buettner, M. Piatek, A. Krishnamurthy, and S. Shenker, "Caching doesn't improve mobile web performance (much)," 2016 {USENIX} Annual Technical Conf. ({USENIX}{ATC} 16), USENIX Association, 2016, pp. 159-165.

[6] J. Nejati, and A. Balasubramanian, "An in-depth study of mobile browser performance," Proc. 25th Int. Conf. World Wide Web, International World Wide Web Conferences Steering Committee, USENIX Association, 2016, pp. 1305-1315.

[7] M. Dasari, S. Vargas, A. Bhattacharya, A. Balasubramanian, S. R. Das, and M. Ferdman, "Impact of device performance on mobile Internet QoE," Proc. Internet Measurement Conf. 2018, ACM, 2018, pp. 1-7.

[8] Z. Wang, F. X. Lin, L. Zhong, and M. Chishtie, "How far can client-only solutions go for mobile browser speed?," Proc. 21st Int. Conf. World Wide Web, ACM, 2012, pp. 31-40.

[9] M. Varvello, J. Blackburn, D. Naylor, and K. Papagiannaki, "Eyeorg: A platform for crowdsourcing web quality of experience measurements," Proc. 12th Int. Conf. emerging Networking EXperiments and Technologies, ACM, 2016, pp. 399-412.

[10] Q. Gao, P. Dey, and P. Ahammad, "Perceived performance of top retail webpages in the wild: insights from large-scale crowdsourcing of above-the-fold QoE," Proc. Workshop on QoE-based Analysis and Management of Data Communication Networks, ACM, 2017, pp. 13-18.

[11] C. Kelton, J. Ryoo, A. Balasubramanian, and S. R. Das, "Improving user perceived page load times using gaze," 14th {USENIX} Symp. Networked Systems Design and Implementation ({NSDI} 17), USENIX Association, 2017, pp. 545-559.

[12] P. Wang, M. Varvello, and A. Kuzmanovic, "Kaleidoscope: A crowdsourcing testing tool for web quality of experience," 2019 IEEE 39th Int. Conf. Distributed Computing Systems (ICDCS), IEEE Press, 2019, pp. 1971-1982.

[13] Navigation Timing API, https://www.w3.org/TR/navigation-timing-2/.

[14] Paint Timing API, https://w3c.github.io/paint-timing/.

[15] Lighthouse, https://developers.google.com/web/tools/lighthouse.

[16] E. Bocchi, L. De Cicco, and D. Rossi, "Measuring the quality of experience of web users," ACM SIGCOMM Computer Communication Review, vol. 46, 2016, pp. 8-13.

[17] RUM-SpeedIndex, https://github.com/WPO-Foundation/RUM-SpeedIndex.

[18] 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," Int. Conf. Passive and Active Network Measurement, Springer, 2018, pp. 31-43.

[19] R. Netravali, V. Nathan, J. Mickens, and H. Balakrishnan, "Vesper: measuring time-to-interactivity for web pages," 15th {USENIX} Symp. Networked Systems Design and Implementation ({NSDI} 18), USENIX Association, 2018, pp. 217-231.

[20] E. Bocchi, L. De Cicco, M. Mellia, and D. Rossi, "The web, the users, and the MOS: Influence of HTTP/2 on user experience," Int. Conf. Passive and Active Network Measurement, Springer, 2017, pp. 47-59.

[21] Kiwi Browser, https://kiwibrowser.com/.

[22] SingleFile, https://github.com/gildas-lormeau/SingleFile.

[23] S. Seo, 2006. A review and comparison of methods for detecting outliers in univariate data sets, University of Pittsburgh.

[24] A. Nikravesh, H. Yao, S. Xu, D. Choffnes, and Z. M. Mao, "Mobilyzer: An open platform for controllable mobile network measurements," Proc. 13th Ann. Int. Conf. Mobile Systems, Applications, and Services, ACM, 2015, pp. 389-404.

[25] B. Jun, F. E. Bustamante, S. Y. Whang, and Z. S. Bischof, "AMP up your mobile web experience: Characterizing the impact of Google's Accelerated Mobile Project," 25th Ann. Int. Conf. Mobile Computing and Networking, ACM, 2019, 3300137, pp. 1-14.

[26] 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," World Wide Web Conf., ACM, 2019, pp. 1532-1543.

[27] S. Egger, P. Reichl, T. Hoßfeld, and R. Schatz, "'Time is bandwidth'? Narrowing the gap between subjective time perception and quality of experience," 2012 IEEE Int. Conf. Communications (ICC), IEEE Press, 2012, pp. 1325-1330.

[28] F. Salutari, D. Da Hora, G. Dubuc, and D. Rossi, "A large-scale study of Wikipedia users' quality of experience," World Wide Web Conf., ACM, 2019, pp. 3194-3200.

[29] Get Started with Analyzing Runtime Performance. https://developers.google.com/web/tools/chrome-devtools/evaluate-performance.

[30] Octane. https://chromium.github.io/octane/.