SUBJECTIVE QUALITY OF WEBPAGE LOADING: THE IMPACT OF DELAYED AND MISSING ELEMENTS ON QUALITY RATINGS AND TASK COMPLETION TIME

Dennis Guse[1], *Sebastian Schuck*[1], *Oliver Hohlfeld*[2], *Alexander Raake*[3], *Sebastian Möller*[1]

[1] Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin, Germany
[2] Communication and Distributed Systems, RWTH Aachen University, Germany
[3] Assessment of IP-based Applications, Telekom Innovation Laboratories, TU Berlin, Germany
Dennis.Guse@telekom.de, Sebastian.Schuck@campus.tu-berlin.de,
Oliver.Hohlfeld@comsys.rwth-aachen.de, Alexander.Raake@telekom.de,
Sebastian.Moeller@telekom.de

ABSTRACT

Webpage loading delays impact the quality perception, may cause frustration as well as annoyance, and can impact the users' behavior. In addition, the quality perception can be further degraded by failures to load webpage components or, in the worst case, the failure to load the entire webpage. This impact depends on the actual implementation of the webpage and will in some cases not only delay the user's task fulfillment, but also makes the task more complicated to solve.

This paper presents the first empirical user study assessing the impact of failed to load elements in comparison to the delayed loading of webpages. The perceived quality is assessed by subjective quality ratings and the *Task Completion Time* (TCT). The results show that *Page Load Time* (PLT) and TCT alone are insufficient quality predictors when considering (partial) load failures. Therefore, established Web-QoE prediction models can be improved by also accounting for load failures.

Index Terms— Web-QoE, Picture Selection, Task Completion Time, Failure to load Elements

1. INTRODUCTION

For *user interfaces* (UI) delay is an important factor contributing to the overall quality perception of the system. Since delay prolongs the *Task Completion Time* (TCT), it should be as low as possible to limit interruptions of the user's workflow [1]. In case that (longer) delays cannot be avoided, providing feedback by showing progress indicators was found to increase the tolerable delay and thus to improve the overall quality perception [1].

For *webpages* delay is a more complex factor since web page rendering involves UI code to be fetched from (remote) servers via a (potentially error-prone) network. To estimate the webpage loading delay, not only the processing delay at both, the server(s)- and the client-side, but also network transmission delays must be considered. In practice, sources of delay are hard to estimate due to varying network traffic conditions, a variety of end-user systems and browser configurations, as well as different server infrastructures.

The loading process of a webpage impacts the user perceived *Quality of Experience* (QoE). QoE is defined as "the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and / or enjoyment of the application or service in the light of the user's personality and current state" [2]. Prior work on perceived quality of loading processes of webpages (Web-QoE) focused mainly on the impact of delay. It can be shown that increasing delay reduces subjective quality ratings (see e.g., [3–7]) and can further result in revenue declines (see e.g., [8, 9]).

The *Page Load Time* (PLT) is the time between the load request issued by a user and the reception of the last byte by the user's webbrowser to fulfill the request. The PLT as *technical measure* can be directly derived from client-side measurements. The user perception can, however, differ from this technical measure. For example, a user can already start solving a task while the webpage is still being fetched, but already partially rendered. This is denoted as *subjective PLT* (sPLT) [4, 5]. Those parts of a webpage that are needed fulfill a task are denoted as *relevant elements* [4].

From a technical perspective, a webpage is a HTML file containing content and references to additional resources including further HTML files, media files, *JavaScript* (JS), and *Cascading Style Sheets* (CSS). JS denotes source code that is executed in a webbrowser. CSS defines the visual appearance of the webpage including color and layout.

Delayed loading and also delayed load failures of referenced resources by a webpage may impact the sPLT and thus degrade Web-QoE. This effect is pronounced since todays' webpages are compiled out of a large body of embedded ob-

jects that are often retrieved from multiple servers [10]. Each server potentially adds delay and thus prolongs the webpage loading process. An unavailable resource can result in an immediate error report to the user or in a timeout, where the later delays the load process even further. Such an impact has, however, not been studied in related work.

This paper complements related work by providing the first study assessing the QoE impact of partly loading webpages (i.e., missing elements such as images or CSS) in direct comparison to the delayed loading webpages. In addition to the subjective quality assessment, the performance of PLT and TCT as common quality indicator is investigated. The results highlight that PLT and TCT alone are insufficient quality predictors in the presence of load failures. This finding motivates the extension of existing Web-QoE models (e.g., [11]) to additionally account for partial load failures.

2. RELATED WORK

The web browsing experience is mainly impacted by two *known* classes of quality indicators [6]. One class resembles visual properties such as the time of the first visual sign of progress [12]. The second class relates to page loading delays and loading pattern, which is in the scope of this paper.

Early indications for web browsing experience were provided by HCI studies on tolerable delays for user interface interaction. These found users to tolerate short delays of up to 0.1 sec for immediate feedback, delays of up to 1 sec for maintaining flow, and delays of up to 10 sec for keeping users focused on the current task [1].

Recent work suggests a logarithmic relationship between subjective quality and PLT [3, 6, 13] that is reflected in ITU-T Recommendation G.1030 [11]. The G.1030 Web-QoE model focuses on information retrieval scenarios like web search.

Additional factors have been found meanwhile that are not captures by the G.1030 model. Strohmeier et al. [4] showed that task-driven interaction is judged more critical than task-free interaction. This observation introduced the notion of *relevant elements* to Web-QoE assessment, i.e., elements needed by a user to solve a given task. In addition, it was found that the temporal position of a relevant element influences QoE, leading to better quality ratings when loaded earlier. For task-driven usage the TCT outperforms PLT for quality prediction [4], which could also be shown for multiple consecutive page requests [5]. In an extension, Guse et al. [14] found situational factors to have no significant impact on quality judgment (e.g., driving in public transport vs. laboratory setting) whereas distraction using a parallel task leads to slightly better judgments.

Web-QoE research is complemented by work in psychology studying acceptable webpage loading delays [15]. This study focused on the impact of a progress indicator on the amount of time until users aborted the load process and reloaded the webpage. It was found that the average waiting time without progress indicator decreased from 13 sec to 3.3 sec from the first to the third non-loading webpage whereas with progress indicator subjects waited first 38 sec and on the third encounter 6.7 sec.

State-of-the-art Web-QoE research is based on the assumption that webpages are only loading delayed. Hence, an understanding of (partial) load failures on Web-QoE is still missing. This paper addresses this gap by presenting the first study assessing partially loading webpages. Delay only conditions are further included to allow this study to be viewed in context of related work.

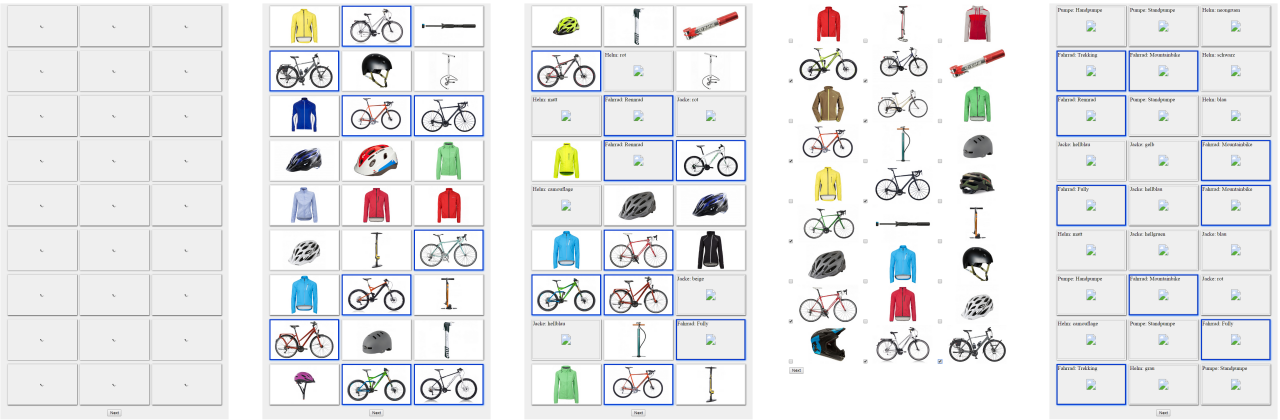
3. EMPIRICAL STUDY

This study assesses the quality of delayed loading and partly not loading webpages. Its design follows a state-of-the-art Web-QoE assessment methodology in which a task-driven interaction scheme is applied. This choice is motivated by the observation that a task-driven interaction is judged more critically than a task-free evaluation (see [4] and [16]). Further, a task that that relies on external resources—which may fail to load but are required to solve the task (relevant elements)—is chosen according to the known relevance of relevant element(s) on Web-QoE (see [5]).

Therefore, the study is based on an image selection task in which a defined subset of images has to be selected. Thus the complete set of images are relevant elements in this task. The chosen task resembles a common scenario in online shopping, where a list of available products is presented (each by an image) and a subset of interesting products is selected by the user. While the impact of partial load failures is assessed in this study, it is ensured that the selection task always remains solvable. This avoids effects due to user frustration, which is not yet completely understood for Web-QoE.

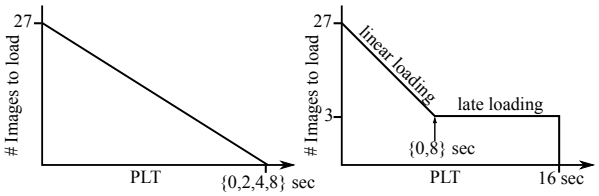
To allow the task to remain solvable, two cases of partial load failures are investigated in this study: *i)* the failure to load task relevant images and *ii)* missing CSS impairing the visual appeal of the webpage. In the case of missing images, task solvability is ensured by providing a textual description that is displayed for each missing image. In contrast to the images, CSS is not required to solve the image selection task and is thus not a relevant element. CSS, however, defines the layout and design of a webpage and thus the failure to load CSS might make solving the task more complicated. In addition, the study of delayed loading without load failures aims to enable the comparison of the partial load failures with prior work on Web-QoE (e.g., [4–6, 11]).

The applied task is to correctly select a subset of images belonging to a particular category out of a total set of 27 images and to press a button when completed. It is then verified that the selection is correct and if not, the user is otherwise asked to correct the selection. For the task to remain solvable in the presence of non-loading elements, textual descriptions of missing images are provided, following the HTML spec-



(a) Throbber (b) Complete selection (c) 1/3 of images missing (d) Missing CSS (e) All images missing

Fig. 1. The webpage for the image selection task. From left to right: All images loading, all images loaded, failure to load images and failure to load CSS. The last four show the solved task.



(a) **Linear** Loading (b) **Late** Loading

Fig. 2. Visualization of the applied webpage load pattern. *Linear*: The loading process of all 27 images is linearly spread over *PLT* time. *Loading*: 24 images are loaded following a linear loading pattern within 0 sec or 8 sec and three randomly selected images load at 16 sec.

ifications [17]. While loading images are represented with a throbber and if the load failed the default error image is shown.

Three different load pattern are applied in this study denoted as *instant*, *linear* and *late* (see Figure 2). *Instant* denotes the references condition, in which the load process finishes immediately, i.e., a PLT of 0 sec. In case of *linear* loading, images are linearly/sequentially loaded from top to bottom where the temporal positions are equally distributed over the specified PLT following [4, 5, 14] (see Figure 2(a)). *Late* loading delays three images until 16 sec and either loads all other images instantly or applies *linear* loading with a PLT of 8 sec to those images (see Figure 2(b)). In the all three load pattern the main HTML file is loaded immediately when starting the task.

This study involves the following classes of degradations:

Page Loading Delay The webpage including CSS is loaded instantly whereas the loading of images is delayed. The PLTs are set to 0 sec (instant loading, following [15] as reference condition), 2 sec, 4 sec, and 8 sec. The loading of all images is linearly spread over the configured PLT (linear loading, see Figure 2(a)).

Missing CSS The webpage and all images load instantly, but CSS fails to load. As a result, checkboxes are presented instead of using a colored frame around a selected image (see Figure 1(d)).

Delayed Images 3 out of 27 images are delayed (late loading, see Figure 2(b)) and will be loaded after 16 sec. Two conditions for load process of the other 24 images are applied: instant loading and linear loading over 8 sec.

Missing Images The complete webpage is loaded immediately (instant loading), but one third (9) or all (27) images fail to load and the alternative text is shown (see Figure 1(e)).

Delayed Missing Images 3 images are failing to load after 16 sec and all images are loading instantly.

The implemented webpage is shown in Figure 1, including the failure to load cases. The task was to select all 9 shown bicycles out of 27 images presenting either bicycles or other bicycle equipment. For the image selection task a webpage was implemented that can simulate delayed loading and failure to load in the webbrowser without requiring a network setup. This webpage was integrated into TheFragebogen[1], which is a HTML5-based JS framework for the implementation of questionnaires. The study was conducted using a Microsoft Surface Pro that provides 10.6" full-HD touchscreen and the digitizer pen was used for input.

[1] `http://www.TheFragebogen.de`

Fig. 3. Continuous 7-point scale (labels in German); labels from left-to-right: extremely bad (0), bad (1), poor (2), fair (3), good (4), excellent (5) and ideal (6) [18].

After fulfillment of each task the subject rated the quality on the 7-point continuous scale [18] shown in Figure 3. This scale was chosen over a standard 5-point absolute category rating scale to enable more fine grained ratings. The rating scale was presented with the following German statement "Bitte beurteilen Sie das Ladeverhalten hinsichtlich der Bearbeitung der Aufgabe", which translates to "Please judge the loading process with regard to the task fulfillment".

In addition to the quality judgment also the TCT is measured by taking the time between the start of the webpage loading and the time the task is solved successfully as indicated by pressing the button with the label "next". Further, the number of unsuccessful attempts to finish a task (clicks on the button) is measured.

The study was conducted in a soundproof cabin according to ITU-T P.800 [19] to avoid impact of environmental distraction. The study involved a training phase in which five stimuli were presented to show the range of impairments and a study phase with one hundred stimuli with each condition repeated 5 times. During this training phase subjects were left alone in the room. In the study participated 28 subjects (11 male and 17 female) aging from 18 to 35 (mean 26). The study was conducted in Berlin, Germany in September and October 2014.

4. RESULTS

This section first discusses the results of the subjective quality assessment followed by the evaluation of the TCT as objective measure. The obtained quality ratings are expressed as *Mean Opinion Score* (MOS), ranging from 0 to 6 (see Figure 3). The TCT is expressed as average by condition (mean TCT). The results are shown in Table 1 for the studied classes of degradations, where each line represents one condition.

4.1. Subjective Quality

A statistical analysis of the quality judgments using Kruskal-Wallis rank sum test showed significant group differences, when comparing all conditions ($\chi(9) = 884.7$, $p < .0001$). A post-hoc analysis with a pairwise Wilcoxon rank sum test with Holm's adjustment for multiple comparison was conducted. The groups of non-significant different conditions are shown in Table 1. That is, conditions represented by the same group number represent non-statistically different results.

Considering only the *Page Loading Delay* degradation, the results show that increasing the PLT from 0 sec to 8 sec decreases the MOS from 5.5 to 3.2. Except for the conditions with a PLT of 2 sec and 4 sec, PLT changes yield significant different quality judgments. This is in line with prior work that investigated load delay alone [4–6, 13, 14].

The *Missing CSS* degradation, which impaired the visual representation of selected items, is *i)* perceived by the subjects and *ii)* resulted in a slightly degraded MOS of 4.1 (cf. 5.5 for the reference condition). In fact, this condition lead to the highest standard deviation (1.4) of all studied conditions. The statistical analysis found the missing CSS degradation to be statistically significantly different to the reference condition (MOS 5.5). Comparing the missing CSS degradation to the delayed loading of images yielded quality ratings that are not statistically different for the *Page Loading Delay* degradation with a PLT of 2 sec and of 4 sec. Thus, those are considered as similar (i.e., forming rating group 2).

The two conditions of *Delayed Images* lead to a statistically not significant different MOS of 2.2 (for 8 sec linear loading) and, respectively, 2.3 (initial loading). Thus, no difference could be found, if all except the 3 images are loaded instantly or linearly.

The *Missing Images* degradation, where the webpage loads except that either $1/3$ or all images fail to load immediately, results in significant different MOS of 2.1 and 1.2. The condition with 27 missing images and an effective PLT of 0 sec yielded the lowest quality judgment of all studied conditions. This degradation is perceived worse than the lowest rated delay only condition that is *Delayed Images*, which had a PLT of 16 sec. In fact, the $1/3$ condition is statistically not significant different from the two *Delayed Image* conditions. Those three conditions are denoted as group 4. The *Delayed Missing Images* degradation, where 3 images are failing to load after 16 sec results in a MOS of 1.7.

4.2. Task Completion Time

For the TCT a Kruskal-Wallis test ($\chi(9) = 959.0$, $p < .0001$) showed significant group differences for all conditions and a pairwise Wilcoxon rank sum test with Holm's adjustment for multiple comparison was conducted. The groups of non-significant different conditions are again shown as rating groups in Table 1.

For the *Page Loading Delay* degradation the TCT increased from 7.6 sec—for an immediately loading webpage—up to 10.8 sec for a PLT of 8 sec. Here, increases in PLT seem to be linearly reflected in the TCT. Similar to the quality judgments the conditions with 2 sec and 4 sec are not statistically different and thus denoted as group 2.

The *Missing CSS* degradation lead to a TCT of 10.8 sec. This is statistically different to the TCT observed for PLTs of 0 sec (non-delayed reference condition), and 2 sec and 4 sec (delayed loading). This condition, however, is not signifi-

Degradation class	Parameters				Quality Rating		TCT	
	PLT	Load pattern	# Images loaded at PLT: loaded	failed	MOS	Group	Average	Group
Page Loading Delay	0 sec	Instant	-	-	5.5	1	7.6 sec	1
	2 sec	Linear	-	-	4.4	*2*	8.4 sec	*2*
	4 sec	Linear	-	-	4.0	*2*	8.7 sec	*2*
	8 sec	Linear	-	-	3.2	3	10.8 sec	*3*
Missing CSS	0 sec	Linear	-	-	4.1	*2*	10.8 sec	*3*
Delayed Images	16 sec	Instant	3	-	2.3	*4*	20.4 sec	*4*
	16 sec	Late	3	-	2.2	*4*	19.5 sec	*4*
Delayed Missing Images	16 sec	Late	-	3	1.7	*5*	20.3 sec	*4*
Missing Images	0 sec	Instant	-	9	2.1	4	14.0 sec	5
	0 sec	Instant	-	27	1.2	6	17.1 sec	6

Table 1. The subjective quality (ranging from 0 to 6; confer Figure 3) and TCT results for all studied degradations. For each the average and statistical non-different groups as found with pairwise Wilcoxon rank sum test with Holm's adjustment is presented.

cantly different to a *Page Loading Delay* of 8 sec (expressed as group 3).

The *Missing Images* degradation, although the webpages is loaded instantly, leads to a TCT of 14.0 sec ($1/3$ of the images missing), and to 17.1 sec (all images missing). This indicates that the selection task becomes more complex to solve when the missing images are replaced by a textual description.

The *Delayed Images* degradations, which both follow the late load pattern and thus have a PLT of 16 sec, lead to a TCT of 20.4 sec with instant loading and of 19.5 sec with linear load over 8 sec for the 24 images. Both conditions are not statistically different and are also similar to the *Delayed Missing Images* condition that resulted in a TCT of 20.3 sec (group 4).

4.3. Discussion

The subjective quality results of the *Page Loading Delay* degradation alone only allow for limited comparability with prior work [4–6, 13, 14]. An absolute comparison is left for future work since different webpages, tasks, load strategies, loading times as well as ratings scales were applied in related studies. However, as found in related studies, the impact of increasing the overall load duration on quality judgments has also been found in this study. This shows that the used image selection leads to similar results with regard to delayed loading alone. However, with regard to load failures of parts of the webpages a comparison to prior work can only be performed indirectly using conditions with delayed loading.

This study shows that all degradations involving load failures lead to a decrease in subjective quality judgments, even when the webpage is loaded instantly and thus those degradations negatively influence Web-QoE. In fact, the results show that load failures are in all cases (except for *Missing CSS*) perceived worse than the delayed loading, even when a high PLT of as much as 16 sec is considered. Despite of the minor visual impact to keep the task always solvable, the case of *Missing CSS* changed the visual feedback of the selected images and thus actually changed the behavior of the UI. It is thus not directly comparable to the case of missing images.

For all failure to load degradations, the results further show that the TCT increases significantly compared to the reference case that instantly loads all images. This leads to the assumption that solving the task correctly becomes more complicated, if the UI of a webpage cannot be loaded completely. The presented study does not allow to conclude, if either the limited UI or the higher effort to solve the task negatively influence the quality judgment.

Similar to Strohmeier et al. [4] it was found that the PLT alone is not a good quality predictor. However, in contrast to related work that suggests to use the TCT instead, this study shows that the TCT is a limited quality predictor in the case of load failures (e.g., for the *Missing Images*, the *Delayed Missing*, and also for the *Missing CSS* case). Therefore, Web-QoE modeling can be improved by accounting for load failures as such degradations present a severe impairment and are not covered by the established indicators PLT and TCT. A modelling approach must be left for future work as the number of studied conditions is rather low and would thus limit the generality of the model.

5. CONCLUSION

In this paper an image selection task was applied to study Web-QoE in the presence of delayed loading as well as failures to load external referenced objects of the webpage. For delayed loading alone the results are in line with prior work. Complementing related work, it could be shown that the failure to load elements—although the task remains solvable—lead to severely lower quality ratings. The results highlight that, although widely used, the PLT and even the TCT alone

do not suffice to predict the Web-QoE in the presence of load failures. The presented study thus aims to pave the way for developing improved Web-QoE models.

Future work is necessary to deepen the understanding of load failures beyond the limited number of conditions investigated in this study. One line of extensions resembles having a task to retrieve several consecutive webpages (session-based usage) as investigated by Strohmeier et al. [5] or ITU-T Rec. G.1030 [11]. Further, proposed extensions should not only focus on task fulfillment in general, but also consider non-ability to fulfill tasks due to load failures. An approach to integrate this was proposed by Leon-Garcia et al. [20] using a session-based approach that also includes non-accessibility and non-retainability as aspects of quality.

6. ACKNOWLEDGEMENTS

Oliver Hohlfeld was supported by the German Research Foundation (DFG) within the Collaborative Research Center (CRC) 1053 – MAKI.

7. REFERENCES

[1] Jakob Nielsen, *Usability Engineering*, Morgan Kaufmann Publishers Inc., 1993.

[2] Patrick Le Callet, Sebastian Möller, and Andrew Perkis, Eds., *Qualinet white paper on definitions of quality of experience Version 1.2*, Qualinet, 2013.

[3] Sebastian Egger, Peter Reichl, Tobias Hoßfeld, and Raimund Schatz, "Time is bandwidth"? narrowing the gap between subjective time perception and Quality of Experience," in *IEEE ICC*, 2012.

[4] Dominik Strohmeier, Satu Jumisko-Pyykko, and Alexander Raake, "Toward task-dependent evaluation of web-QoE: Free exploration vs. "who ate what?"," in *IEEE Globecom Workshops*, 2012.

[5] Dominik Strohmeier, M. Mikkola, and Alexander Raake, "The importance of task completion times for modeling web-QoE of consecutive web page requests," in *IEEE QoMEX*, 2013.

[6] Sebastian Egger, Tobias Hoßfeld, Raimund Schatz, and Markus Fiedler, "Tutorial: Waiting Times in Quality of Experience for Web based Services," in *IEEE QoMEX*, 2012.

[7] Sebastian Egger, Tobias Hoßfeld, Raimund Schatz, and Markus Fiedler, "Waiting times in quality of experience for web based services," in *IEEE QoMEX*, 2012.

[8] Eric Schurman and Jake Brutlag, "Performance related changes and their user impact," in *Velocity*, 2009.

[9] Greg Linden, "Make data useful," 2006.

[10] Michael Butkiewicz, Harsha V. Madhyastha, and Vyas Sekar, "Understanding website complexity: measurements, metrics, and implications," in *ACM IMC*, 2011.

[11] International Telecommunications Union, "G.1030: Estimating end-to-end performance in IP networks for data applications (02/2014)," 2014.

[12] Heng Cui and Ernst Biersack, "Trouble shooting interactive web sessions in a home environment," in *ACM SIGCOMM Workshop on Home Networks*, 2011.

[13] Dennis F. Galletta, Raymond Henry, Scott McCoy, and Peter Polak, "Web site delays: How tolerant are users?," *Journal of the Association for Information Systems*, vol. 5, no. 1, 2004.

[14] Dennis Guse, Sebastian Egger, Alexander Raake, and Sebastian Möller, "Web-QoE under real-world distractions: Two test cases," in *IEEE QoMEX*, 2014.

[15] Fiona Fui-Hoon Nah, "A study on tolerable waiting time: how long are web users willing to wait?," *Behaviour & Information Technology*, vol. 23, no. 3, pp. 153–163, 2004.

[16] International Telecommunications Union, "P.1501 subjective testing methodology for web browsing (02/2014)," 2014.

[17] W3C, "Recommendation: HTML 4.01 Specification," 1999.

[18] International Telecommunications Union, "P.851: Subjective quality evaluation of telephone services based on spoken dialogue systems (11/2003)," 2003.

[19] International Telecommunications Union, "P.800: Methods for subjective determination of transmission quality (08/1996)," 1996.

[20] Alberto Leon-Garcia and Leon Zuchermann, "Generalizing MOS to assess technical quality for end-to-end telecom session," in *IEEE Globecom*, 2014.