On the Experimental Biases in User Behavior and QoE Assessment in the Lab

Werner Robitza*, Péter A. Kara†, Maria G. Martini†, Alexander Raake‡
*Telekom Innovation Laboratories, Deutsche Telekom AG, Berlin, Germany
Email: werner.robitza@telekom.de
†WMN Research Group, Kingston University, London, UK
Email: {p.kara, m.martini}@kingston.ac.uk
‡Audiovisual Technology Group, Technische Universität Ilmenau, Ilmenau, Germany
Email: alexander.raake@tu-ilmenau.de

***Abstract*—User behavior is one of the key components of customer engagement and abandonment, which result from a good or bad Quality of Experience. However, methods to evoke and measure user behavior are still understudied. This paper presents an in-depth look at a study in which we measured user behavior during video streaming consumption in a controlled laboratory environment. We confronted subjects with typical streaming problems such as stalling and quality fluctuations. The subjects were not informed about the real purpose of the test; their behavior was tracked unobtrusively. The results suggest that the method can elicit responses to the inserted problems, such as seeking, pausing, or reloading the web page. However, a third of the subjects acted apprehensively, meaning that they changed their behavior due to being part of a test. In this contribution, we elaborate on the underlying reasons for those experimental biases. We discuss the suitability of different test designs for behavioral assessment and give guidelines on how to quantify and combat biasing factors introduced by the test procedure.**

I. Introduction and Motivation

The ever-increasing demand for bandwidth in Internet video services often results in problems for users such as long loading times or low video quality (e.g., in the case of HTTP adaptive streaming), which in turn may lead to users abandoning a service if they are not satisfied [1]. Over-the-Top providers (OTT) therefore aim at increasing user *engagement*, in particular the time spent using their service. Ultimately, engagement is a sign of good Quality of Experience (QoE)—abandonment is a lack thereof. Current subjective test methodologies (e.g., from ITU-T Rec. P.910) cannot be used to assess specific user actions that are indicators of bad experience, such as cancelling a video session. To obtain insight into the users' motivations behind a certain action, we therefore have to find new methods to study these aspects in a laboratory setting. Furthermore, in the future, instrumental QoE models may not only predict Mean Opinion Scores, but engagement or abandonment—based on the results of such tests.

This paper takes a different perspective on a "behavioral QoE" test design we already presented in [2]. We studied user behavior as a response to quality problems (including video stallings and resolution/bitrate changes). We wanted to find out whether it was possible to elicit user reactions in a lab context, using a "deceptive" study (with a so-called "mock task"): users were initially *not* told that their behavior would be monitored. They did not know that the problems were inserted on purpose and found out about it only after the test. As a mock task, subjects were asked to describe the video contents. While we saw specific user behavior as a response to those problems (e.g., seeking forward/backward when a stalling event occurred), we discovered that a fourth of our participants reacted *apprehensively*. This means that they changed the way they behaved, due to the mere fact that they were taking part in a test. For example, some users did not want to reload the page out of fear to "destroy experiment data"—clearly an experimental bias.

Why do these biases exist, how can we measure and avoid them? We address these questions by presenting an extended version of our previous study, with more subjects and a much more in-depth look at how users behave. The main purpose is to investigate the so-called *demand characteristics*, which we will describe in Section II. Which factors influence users in such test settings? How can their effects be limited in order to make the study more ecologically valid? Our study setup is briefly described in Section III. In the results (Section IV), we focus on the participants, their characteristics, and their (behavioral) responses to quality problems. We discuss the impact of those factors in Section V, followed by conclusions in Section VI.

II. Related Work and Psychological Background

The idea of tracking user actions in a laboratory context, related to QoE, is not per se new. Mok *et al.* [3] presented a study in which they exposed subjects to different network conditions while consuming video streamed over a local network. The participants' behavior was monitored and related to certain QoE characteristics. However, the authors did not describe the specific tasks and experimental protocol, which leads to questions concerning the reproducibility and experimental validity of the study. Were subjects behaving in a natural way,

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 643072, Network QoE-Net.

or were they influenced by knowing that their behavior was tracked?

For behavioral psychologists, *demand characteristics* (or *demand effects*) are well-known concepts. Orne [4] explains how and why subjects change their behavior because of their participation in an experiment, their assumptions about the experiment purpose, the experiment's setting, the task, etc. As a consequence, participants may want to act in a manner that results in a good impression of themselves or confirms the underlying study hypothesis. This could happen both consciously and unconsciously. Notably, demand characteristics are different from *experimenter biases*, in which the presence and acting of an experiment leader influences the results. In fact, we may conceive a study in which no human experimenter is present but which still carries intrinsic demand characteristics, such as in tests performed remotely (e.g., crowdsourcing [5]).

In engineering, this perspective is often neglected; humans are sometimes expected to deliver responses in an almost mechanical, neutral way to the stimuli to which we expose them. This approach is considered reasonable for quality assessment tests or comparative studies, where, for instance, participants may not feel compelled to give positive ratings. However, as we progress towards a more holistic view on QoE—not just looking at "visual quality" but, amongst other factors, the context of use—demand characteristics will play a larger role. The reason is that we would like to use more realistic services for testing, in more (ecologically) valid contexts, with subjects being more involved and immersed, which makes room for subjective and personal decisions, based on more complex cognitive processes. These processes may be more influenced by experimental biases.

III. Experiment Description

The main procedure of our experiment has already been described in [2]; therefore, in this section, we only briefly summarize the main parameters and focus more closely on the protocol itself.

1) Source Material: We selected 32 video clips from large online video portals, selecting various genres in order to fit users' preferences. All of them were cut to a length between 1.5 and 3 minutes, if necessary, to show an interesting portion. We then prepared them for adaptive streaming by encoding them into five quality representations, from 240p resolution (12 fps, 150 kBit/s video, 32 kBit/s audio bitrate) at the lowest profile, up to 1080p (24 fps, 4.5 MBit/s video, 128 kBit/s audio bitrate) at the highest profile. Therefore, the profiles spanned a large quality range.

2) Instructions and Tasks: The subjects were instructed (in written form) to select a video of their choice from an overview grid, displayed on a customized website in a web browser (Google Chrome). Every selected video was watched on a separate web page, thus, we simulated a typical usage pattern from video on demand websites (e.g., YouTube). Once the video finished, participants were asked to describe the video contents in a few sentences, as if they were sharing it with a friend via social media. Also, they had to answer a question about the content (e.g., "What was the color of the main actor's shirt?"), to prove that they had paid attention. Finally, they could rate how much they liked the video clip on a 5-star scale. This procedure would repeat seven times, every time with a new, freely chosen video.

3) Conditions: For each video the participants selected, a different random playback condition was chosen. The first video was always the reference, that is, it started immediately and played fluently. This was done to give people the impression that the service was working well. Then, for every subsequent video, a new error condition was selected randomly, simulating network outages. The chosen error conditions were: 1) 30 seconds initial loading time, 2) 30 s of stalling inserted at 00:30, 3) quality drop from highest to lowest resolution at 00:45, 4) constant medium quality level (480p), and 5) constant lowest quality (240p). Note that stalling was indicated to the user with a "loading dots" animation. Finally, the reference condition was shown again. During playback, all interactions with the video player and the web page were monitored in the background using JavaScript. The actual video viewing session lasted about 25 minutes.

4) General Protocol: The test started with a written introduction on its purpose and a general questionnaire on online video usage. Then, the experimenter gave the test device to the participant, a 13" MacBook Pro Retina. The subject could sit on a sofa in a living room-like environment, having been told to imagine a viewing situation at home. During the main part, the experimenter sat outside the test room, pretending to be busy. In reality, he was waiting for subjects to react to problems, for instance by calling for help when the videos would not load.

After the main part of the test, the experimenter asked the subject, "Did everything go well?", then listened to the verbal descriptions of playback issues—if the user actually noticed any. If a subject hinted at problems, the experimenter would act surprised and ask about details on what had happened. However, when the participant did not mention any problems on their own, they were asked whether they had noticed that some clips would not load, or were played with "bad quality". Only then, the real purpose of the test was revealed, that is, studying the user reactions to playback issues. A discussion phase followed, in which we assessed several points: 1) what participants had done in the case of problems (or whether they had not reacted at all), 2) what their typical reaction at home would have been, 3) whether and why they had reacted differently in the laboratory (i.e., the study context) compared to real life. Finally, a longer questionnaire was handed to the subjects, with questions on problems they have experienced in video services and how they typically deal with them.

IV. Results

A. Subject Sample

Overall, 25 subjects took part in the study, 13 of them female. The age range was 19–60 (median 30). The first 15 subjects (whose results were the basis for [2]) were recruited

using a dedicated portal to finding study participants; the remaining 10 were then acquired via public postings on classifieds.

Due to the selection procedure, one major discriminating factor among the pool was the subject *naïvety*, in terms of previous experience with such kinds of experiments. 18 subjects had already participated in studies with computer-based test procedures, 13 of them in a video quality rating test. Only 7 were completely naïve (in the sense of never having taken part in any study before). Note that at this point, we could not identifiy a statistically significant influence of naïvety in terms of user reactions, but we will address this factor in future research.

B. Behavioral Responses

In [2] we listed several behavioral (inter)actions that are typical for video on demand services. Similar reactions are also mentioned in the survey results shown in [1]. In particular, we want to focus on corrective actions, that is, behavior intended to resolve problems that occur during playback. For instance, users may try to click *pause* and wait for the player's buffer to fill again before continuing, since they have learned that it results in a smoother streaming experience. In the following, we look closely at the observed reactions to the two main inserted degradations, stallings and quality drops.

1) Responses to Initial Loading / Stalling Events: For the stalling events (initial loading and stalling during playback), we observed typical reactions from 19 users (76% of the subjects), with differing motivations.

- *Seeking:* Seeking backwards or forwards into the unbuffered region of a video was used by 13 participants. Users may assume that some videos always get stuck at a certain point, and their interaction "pushes" the video over that error mark (as explained by several subjects during the tests).
- *Clicking:* Seven participants were moving the mouse and clicking on non-interactive regions of the player or the window. Of course, this had no effect on the playback.
- *Selecting another video:* Six users tried clicking the browser's *back* button, then selected a video from the grid again. This may be motivated by users thinking that only particular videos exhibit playback problems.
- *Reloading page:* For five users, reloading the page via the browser button was a way to mitigate problems. The underlying rationale was that this would "reset" a connection to the server.
- *Pausing:* Four subjects thought that pausing and letting the video stream buffer would cause it to play more fluently afterwards, or buffer more quickly.

We identified several reasons why users did not react or reacted differently compared to real life. In the discussion phase, some subjects stated multiple of the reasons listed below.

- *Apprehensiveness:* Seven users did not want to influence or manipulate the test process. We will describe those results in detail in Section IV-C.

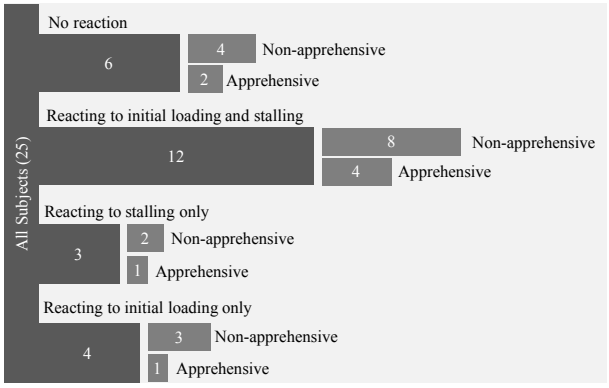

Fig. 1. User reactions to inserted conditions considering apprehensiveness.

- *Low annoyance:* Five users found the stalling events to be not that annoying, for instance because they occurred only once or they were not too long. We did not expect this, as we had thought that 30 seconds of stalling would stronly annoy users. This finding also contrasts with experiences from real-life, where, for example, cancellation rate was found to be as high as 80% for 30 seconds of loading time [6]. Although the context of use (paid service, watching movies vs. casual browsing of short clips) is a significant influencing factor in such considerations, our finding still highlights the strong influence of the test setting on the ecological validity and interpretability of the gathered results.
- *Task dependency:* One user was annoyed by the events, but did not consider them to be detrimental to the given task. In general, users may assume that they will receive a remuneration for the test if they just wait long enough.

2) Responses to Low Quality / Quality Drops: To our surprise, no subject reacted in a (visible) way to the conditions where the quality/resolution was changed in comparison to the reference (including the ones with constant medium/low quality). Based on the interviews, we identified several reasons according to which our participants did not respond to the conditions:

- *Technical ignorance:* Six users did not know if and how they could change the quality—although the player offered a button to change it. Note that this may also depend on the instructions and familiarity with the system under study (i.e., subjects being unfamiliar with the specific player software).
- *Low annoyance:* For six users, the low quality was not annoying enough to cause a reaction. Some explained that they found the videos so entertaining that they did not care.
- *Apprehensiveness:* Similar to the case of stalling, five users wanted to react by interacting, but did not dare to.
- *Task dependency:* Five users found the low quality to be annoying, but did not care to react because it was not relevant for their task performance (i.e., describing the content and answering the content-related question).
- *Imperceptible:* Four users did not recall seeing any video with low quality, despite the inclusion of two streaming conditions that show the lowest 240p quality level.
- *Attribution of problems:* One user mentioned that she

thought the source video was already stored in bad quality. Our intention was to simulate network issues, but the subject clearly did not blame the network.

Looking at the results it becomes apparent that the human perception of quality drastically changes when subjects are not required to rate audiovisual quality. To summarize our most important findings, which we will address in the following, Figure 1 shows the distribution of the user reactions and their apprehensiveness.

C. Subject Apprehensiveness

We have already stated several reasons for users not reacting or reacting differently compared to real life, including low annoyance or task dependency. However, when the reason is that they are concerned about the test outcome, we call them "apprehensive". How can we quantify apprehensiveness in experiments? The easiest way would be to ask subjects—they provided generous verbal explanations about the reasons for their behavior. In other words, subjects were not shy to admit the impact of the experiment context on their conduct. According to their self-descriptions, eight subjects acted apprehensively (i.e., a third of all participants, see also Figure 1). This behavior usually stems from a conscious process. For instance, a subject could be deliberating about the consequences of hitting the *reload* button, then deciding not to press it, since they are not sure of the technical impact on the experimental data.

In our tests, apprehensive subjects mentioned that they feared they would "destroy" test data by interacting with the player or page. Another reason was that no interactions other than selecting videos had been explicitly pointed out in the instructions (e.g., reloading the page). Thus, we can see two underlying rationales for the modified behavior: 1) fear of changing experimental data, 2) wanting to follow instructions to the letter. Those are not necessarily orthogonal, but we may hypothesize that users belonging to those groups show different levels of obedience and care in test situations. In the end, the underlying reason may be that subjects do not want to appear disobedient, or have their results excluded from the study.

It would be preferential to determine the factors that lead to apprehensiveness. Since at this point we still lack a comparative study in other environments (e.g., at home or through crowdsourcing) or with other systems (e.g., using a real video portal instead of a simulated one), those independent variables cannot be considered yet. Consequently, we can only rely on factors present in our sample, in which the number of subjects is too low to infer statistically significant relationships. However, we could observe a tendency for female participants to be less likely to be apprehensive (Fisher's test, $p = 0.097$).

D. Other Demand Characteristics and Biases

In addition to the specific cases of users not wanting to interact with the system (due to the instructions or out of fear to deliver wrong data), we found several examples of behavior that differs from what literature suggests occurs in real life. For instance, some users described that they were not annoyed by the stalling periods to begin with. This is an effect of a more general experimental or contextual bias; in the lab, our subjects appeared to be less critical with regard to degradations. At first sight, this statement sounds counterintuitive, as we know from quality assessment tasks that in fact the opposite is true (see, e.g., [7]). The difference lies in the given (or assumed) task. Only when we instruct subjects to rate the quality, they become more critical.

We assume that the involvement in the test itself will impact this result, determined by, for example, the subjects' reimbursement, the perceived importance of the research, or the familiarity with the hardware and software being used. In order to find out more about those factors, further studies have to be conducted.

V. Discussion

The above results lead us to three major questions, which we will discuss in this section: 1) What can be learned about the use of deceptive studies for behavioral QoE? 2) How do we assess and deal with demand characteristics before, during, and after a test? 3) What are the possible alternatives to lab testing, under the assumption that the demand characteristics are the strongest in this scenario?

A. Debating the Usefulness of Deceptive Studies

It is obvious that demand characteristics are strongly present in behavioral tests, leading to subject apprehensiveness and distorted results. Behavioral psychologists have of course dealt with those issues for decades; they have devised methods to counteract the problems associated with those effects [8], including deceptive studies in which the research question is hidden.

It is easy to imagine that if the purpose of the study was made clear to the subjects, a considerable percentage of them would even act more apprehensively than we have seen. For example, if participants were told that the experiment purpose is to see how they will react to a video stalling event, this could already unconsciously bias their preconceptions. Even more so, they could anticipate stalling events, deliberate about their possible actions, and act in a way they think the experimenter is expecting from them. Such a problem would be exacerbated by instructions telling people which actions are possible; it would appear that those actions are wanted by the experimenter. The results from such a study would therefore be questionable in terms of validity.

At the same time, deceptive studies have strong drawbacks. We can compare the reactions (from Section IV) to what our subjects have already tried in real life (in the case of quality problems), as shown in Figure 2. We can see that interaction with the player itself is the top choice, with other reactions being restarts of software or hardware. Some of those actions could be made possible in a laboratory test—such as choosing another browser or player, or restarting/rebooting equipment—but it would be hard to communicate this possibility to subjects without hinting at the real experiment purpose. In other words,

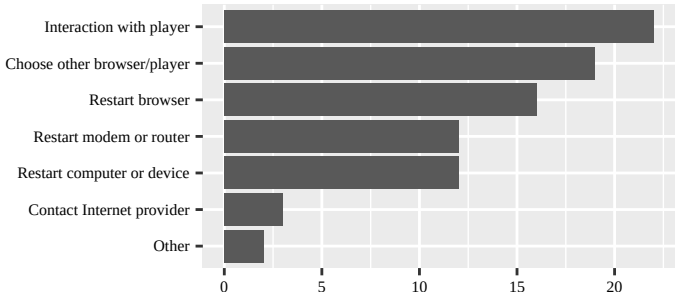

Fig. 2. Count of all subjects' responses to: "What have you already done in the case of quality problems in real life?"

subjects may think, "Why would the experimenter tell me that I can restart the browser?" Certainly, most subjects would not dare to reboot a test computer unless explicitly instructed to. Also, there is no realistic chance to elicit feelings of "I want to file a complaint with my Internet provider" if the situation is as decontextualized as in the lab.

In addition to the above issues, we cannot show as many conditions or stimuli as one typically would include in a classic audiovisual quality rating test. In fact, we walk a thin line between designing an "efficient" test (in the sense of testing as many conditions as possible) and not hinting at the real research questions behind—or at least making the subject suspicious. We can also assume that the typical reaction to QoE problems would change over the course of a (repetitive) behavioral test.

Considering the above, the following question can be raised: in the domain of QoE, how can we still validly assess behavior in the lab, let alone assessing the interplay of QoE and behavior? Subjective quality assessment heavily relies on questionnaires and standardized rating scales, telling users what is expected from them. For example, one would ask subjects to rate the audiovisual quality of video stimuli after they have been trained on a set of stimuli that exhibit the full quality range. It is beyond the scope of this paper to scrutinize this approach in general. From the above discussions however, it should become clear that—once we consider user behavior in relation to quality—it is not enough to simply ask users to provide a quality rating after a stimulus presentation.

Concluding, we still believe that the shown paradigm is useful and versatile in assessing user behavior in the lab. Most importantly, the main part of the study facilitates interesting, profound discussions with the subjects about *why* they acted a certain way, from which more insight into underlying motivations can be obtained. Moreover, the shown method is not only applicable to video streaming, but also provides interesting research opportunities for telephony or IPTV services. Finally, we believe that an extension of current standardized testing methodologies to capture behavioral responses would be useful.

B. Handling Demand Characteristics

Since demand characteristics are necessarily part of any study involving humans, we can only seek to alleviate their

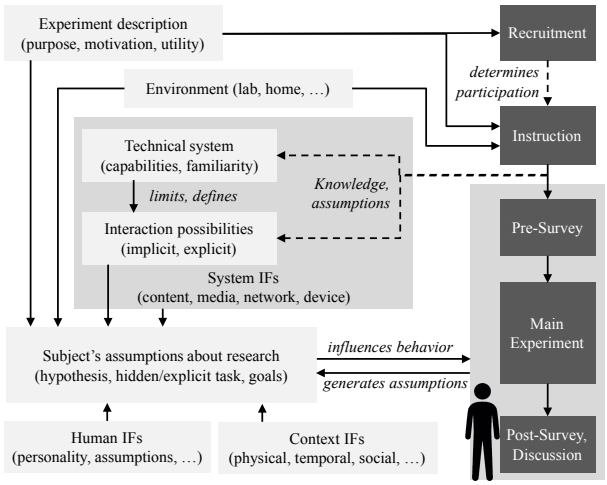

Fig. 3. Experimental influencing factors (IFs) for behavioral QoE research. Dark grey boxes represent the subject's state in the respective process.

effects by proper testing. In practice, many techniques for that have been already proposed, some of them more basic, for example as outlined in Section 11 of ITU-T Rec. P.913. The guidelines relate to the preparation and conduction of tests (including instructions and general procedures) and were also followed in the study presented in this paper, where applicable. However, it should be noted that those recommendations are primarily developed for audiovisual quality rating tasks; fully interactive or behavioral QoE tests have not been considered.

Deceiving participants may be a method of choice, provided that rules of ethical research are followed [9]. Hidden and passive measurements will make participants less aware of the measurement context. Reducing the involvement of the experimenter may also help; any kind of personal influence should be eliminated by using written and automated test instructions. Discussions should stick to a specific protocol.

After a behavioral experiment—no matter if it is using a deceptive paradigm or not—it is crucial to ask participants about what role they thought they played in the study, how they judged their awareness of the research hypothesis, and whether they acted apprehensively. In our survey, we asked subjects whether they at some point became aware of the fact that the test was not (just) about describing video contents. 13 (i.e., half) of them answered "yes" to that question. This sounds like a high percentage, but it may be due to either the larger number of users with prior video quality testing experience, or a certain bias intrinsic to the question itself. Concerning the latter point, we could identify at least three users who in the questionnaire responded that they became aware of the deception, despite having admitted to have had no clue about it while being interviewed. Clearly, one of the two statements must be false. Those subjects are either trying to play "smart" by pretending to know, or play "good" by pretending not to know about the test's real purpose. It is therefore a great challenge to develop methods that neutrally capture the subject awareness of what is being tested.

To facilitate understanding of what influencing factors (IFs)

exist and how they should be taken into consideration, Figure 3 gives a non-exhaustive overview of the different IFs in behavioral QoE experiments. Here, on the right hand side, we assume the subject going through the process of being recruited, instructed, and then exposed to the actual test procedure. Various IFs (including those mentioned in [10]) affect those processes, which in turn generate assumptions about the research. Those assumptions lead to modified behavior during the test stages. Note that the experimenter him-/herself is not included in this figure, although we expect the presence of him/her to also influence the results.

C. Alternatives to Lab Testing

Until now, we have described several impacts of laboratory test situations on user QoE and their behavior. Naturally, the question arises whether a different testing context would yield other results, that is, more natural responses—or even cause more demand effects? A change of context will also change the user's assumptions [10], influencing the way a certain (rating, viewing) task is completed, but also very likely changing the perception of quality and its impact on behavior.

In our lab test we shifted the typical laboratory setting as close as possible to a home-like experience, with the help of a living room-like environment, a relaxed viewing position, the use of a laptop, and the casual task of summarizing videos for friends. Compared to standards-compliant video quality testing procedures, the test can be described as a more ecologically valid lab study. Still—as indicated by the discussion with our subjects—it remains to be seen whether the results are merely artifacts of the context in which the study was carried out.

What other methods apart from laboratory testing exist that could be adopted in our future research?

1) Crowdsourcing: By using dedicated platforms to recruit test participants and conduct the tests, crowdsourcing enables users to participate in experiments from their home, anonymously [5]. It allows for gathering more results than in a lab test, in a shorter amount of time. Also, the impact of the experimenter is less present. However, the challenge lies in detecting unreliable, potentially cheating users. Another drawback is that live discussions with subjects become impossible, and collecting more in-depth feedback is hard to implement.

2) Passive large-scale measurements: The browser plugin *YouSlow* [11] is an example of a software any interested user can install. It tracks YouTube stalling and resolution change events as well as user viewing time, and sends them to a central database. Such an architecture can be used to assess user behavior from an outside perspective, without access to the streaming infrastructure. However, it may not allow detailed insight into users' motivations.

3) Longitudinal studies with friendly users: Combining laboratory-style discussions with long-term real-life service usage, this approach could make use of feedback forms and passive user behavior tracking. Participants could be given test systems or use real services, and their reactions to QoE problems could be investigated on a case-by-case basis. The downside of this approach is that it is very time- and resource-consuming.

VI. Conclusions and Future Work

In this paper, we showed an extended version and different perspective on our pilot test presented in [2]. We observed user reactions to typical video streaming problems such as stalling or quality drops. Our main focus was on the demand effects that exist in psychological experiments, which cause subjects to change their behavior depending on their assumptions about the purpose of the study. It was shown that a third of the participants did not act as they would under everyday viewing conditions. We highlighted factors that lead to user apprehensiveness and discussed the impact of those characteristics on the way tests in the domain of QoE are typically performed. We gave guidelines on how to handle demand characteristics, and then discussed alternatives to the current lab-based approach.

In the future, we will repeat this test paradigm, changing contextual factors such as the environment (e.g., in the user's home, or through crowdsourcing), to see how strongly present demand characteristics are in these situations. A longitudinal study with friendly users may combine behavioral assessment from home with the insight that experimental lab research gives.

References

[1] Conviva, "How Consumers Judge Their Viewing Experience," Conviva, Tech. Rep., 2015.

[2] W. Robitza and A. Raake, "(Re-)Actions Speak Louder Than Words? A Novel Test Method for Tracking User Behavior in Web Video Services," in *Eigth International Workshop on Quality of Multimedia Experience (QoMEX)*, Lisbon, June 2016.

[3] R. K. P. Mok, E. W. W. Chan, X. Luo, and R. K. C. Chang, "Inferring the QoE of HTTP video streaming from user-viewing activities," in *Proceedings of the first ACM SIGCOMM workshop on Measurements up the stack - W-MUST '11*, 2011, p. 31.

[4] M. T. Orne, "Demand characteristics and the concept of quasi-controls," in *Artifacts in Behavioral Research: Robert Rosenthal and Ralph L. Rosnow's Classic Books*, 2009, p. 110.

[5] T. Hossfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "Best practices for QoE crowdtesting: QoE assessment with crowdsourcing," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 541–558, 2014.

[6] S. S. Krishnan and R. K. Sitaraman, "Video stream quality impacts viewer behavior: Inferring causality using quasi-experimental designs," *IEEE/ACM Transactions on Networking*, vol. 21, no. 6, pp. 2001–2014, 2013.

[7] N. Staelens, S. Moens, W. Van den Broeck, I. Marien, B. Vermeulen, P. Lambert, R. Van de Walle, and P. Demeester, "Assessing quality of experience of IPTV and video on demand services in real-life environments," *Broadcasting, IEEE Transactions on*, vol. 56, no. 4, pp. 458–466, 2010.

[8] R. Rosenthal, R. L. Rosnow, and A. E. Kazdin, *Artifacts in Behavioral Research: Robert Rosenthal and Ralph L. Rosnows Classic Books*. Oxford University Press, 2009.

[9] American Psychological Association, "2010 Amendments to the 2002 "Ethical principles of psychologists and code of conduct"." *The American Psychologist*, vol. 65, no. 5, p. 493, 2010.

[10] U. Reiter, K. Brunnström, K. De Moor, M.-C. Larabi, M. Pereira, A. Pinheiro, J. You, and A. Zgank, "Factors influencing Quality of Experience," in *Quality of Experience: Advanced Concepts, Applications and Methods*. Springer, 2014, pp. 55—-72.

[11] H. Nam, K.-h. Kim, and H. Schulzrinne, "QoE Matters More Than QoS: Why People Stop Watching Cat Videos," in *IEEE International Conference on Computer Communications*, April 2016.