# Analysis of Assessment Alteration Phenomena of Subjective Quality of Experience Measurements in 2D and 3D Mobile Video Services

Péter András Kara, László Bokor, Sándor Imre
Mobile Communications and Quantum Technologies Laboratory – Multimedia Networks and Services Laboratory
Department of Networked Systems and Services (HIT), Budapest University of Technology and Economics (BME)
Magyar Tudósok krt. 2, H-1117, Budapest, Hungary
E-mail: {kara, goodzi, imre}@mcl.hu

***Abstract*— The growing importance of Quality of Experience over Quality of Service demands precise results in the monitoring of experienced quality; empirical assessment of subjective QoE measurement on perceived quality is expected to deliver accurate reflection of reality. The goal of this paper is to highlight potential errors in existing subjective QoE measurement methodologies. Our approach focuses on a special topic of distortions caused by preconceptions based on prior technical knowledge of evaluation measurement test subjects. The paper presents two series of measurements where the test subjects were aware of the service parameters during the evaluation of the given services. The paper specifies the identified distortion phenomenon and shows how cognitive dissonance played a role in the formation of evaluation patterns and the distortion of the Mean Opinion Score.**

***Keywords: Quality of Experience, Quality of Service, Mean Opinion Score, 3G HSDPA, 2D and 3D video services, cognitive dissonance***

## I. INTRODUCTION

One of the most important pillars of modern society is the provision and consumption of services. The list of properties of a service provides comparable information to the consumer. Although this does seem to be the universal method of comparison between services of the same kind, it must not be ignored that it is not the equivalent of actual user experience. This means that no matter how high such properties score if the service does not satisfy the consumer. For instance, in case of a video chat which uses mobile Internet connection, it is totally irrelevant how staggering the bandwidth is when the two participants of the conversation have a hard time understanding each other. This leads to the conclusion that the true value of a service rather lies in the "degree of delight or annoyance of the user" [1] (Quality of Experience – QoE) than the "totality of characteristics" [2] (Quality of Service – QoS).

Of course QoE and QoS are unquestionably connected, but their precise relationship is hard to define. However, there are some promising recent researches to flawlessly forge QoE values from a set of QoS parameters (e.g. [3]), yet a widely accepted method is still lacking. Service providers inevitably require user feedback on end-to-end performance to reach a cost effective level of QoE. Monitoring QoE primarily benefits for service providers, but on the other hand, it improves reception for subscribers.

Because of its importance, QoE monitoring is a well defined, standardized process [4]. However, the results of such measurements are affected by environmental information, for example the type of connection, location, device or even some available QoS parameters. In this study, we introduce that the usage of such information depends on the subject's prior technical knowledge and experience on the present technology (Level of Comprehension – LoC). Our term of Level of Comprehension [5] could be defined as "the amount of one's prior technical knowledge and experience which deduces and implies the possible usage of environmental data". In some cases, the awareness of parameters regarding the service cannot be avoided; therefore the results are preordained to be altered. Several examples can be mentioned from everyday life, where the preconceptions create distortions in user experience. The direction and power of these effects are quite far from triviality, yet it hasn't been circumspectly analyzed so far. Mobile video services demand accurate measurements and could benefit from the avoidance or at least the reduction of such distortions.

The complete definition of QoE also states that "it results from the fulfillment of his or her expectations" [1]. In this case, "expectation" refers to the desired level of quality which one has towards a specific service. However, a different interpretation of this word also plays a significant role. The word expectation also means the level of quality one anticipates to experience; a prior idea, a preconception of quality. One could easily presume perceived quality to utterly match these anticipations, but what happens when it doesn't? That would create a disharmonic state between the objective cognition of perception and the subjective cognition of preconception. The theorem of cognitive dissonance [6] explains the different methods of dissonance reduction that could occur in such a situation.

This article deals with the following topic: we study how the combination of aforementioned QoS parameters and different LoC levels alter the assessment results of two different QoE measurements. We also examine the unavoidable psychological reason which empowers preconceptions and manages to alter the refinement of perception. Measurement M1 was the evaluation of a video conference performed on a real-life 3G HSDPA network, while M2 was aiming at 3D multimedia streaming through a GPON transport and Wi-Fi access network. In both cases, the objective of the test participants was to grade the experienced quality, while possessing the parameters of the connection. The research goal was identify the alteration phenomena in both cases and to find correlation between the distinguished levels of LoC and the altered results; how prior knowledge and experience influence QoE.

The article begins with the introduction of assessment alteration approaches with some up-to-date examples and related work in Section II, followed by the configuration and the results of our experiments in Section III and IV. The last section concludes the paper, containing the possible future directions of this topic.

## II. BACKGROUND AND RELATED WORK

The field of subjective determination of transmission quality has well defined standards, intensely detailed recommendations, and countless of exceptional papers sharing the experiences of researches and measurements. The ITU-T P series [7] provide a wide range of recommendations relating to the topic. A fine example for a subjective, context-aware, real-life QoE measurement was conducted by I. Ketykó et al. [8], dealing with the interference of the location and the number of surrounding people on perceived quality. Test subjects in both of our measurements were isolated in a fixed location, yet it is an exciting idea to investigate the explicit effects of the environment. We find it even more interesting to analyze the implicit effects of environmental information in case of varying location, which is a possible continuation of our topic. Explicit effects like the rise of environmental noise level due to the presence of vehicles, machinery or people affect perception and focus without a doubt, but we assume that awareness regarding the environmental information comes with its own distortions. Just for instance, a mobile location like urban public transportation implies mobile access to the service network. Mobility-awareness can create preconceptions and thus shape QoE measurement results, which is the topic of one of our current projects.

Of course the usage environment is not the only factor that affects quality. G. Exarchakos et al [9] highlights how the level of perceptual quality relies on the specific content and network impairments. Although our measurement M1 featured motion in the video conversation (e.g., the test moderator moves from one part of the camera field of view to another), still did not show high vulnerability towards packet loss due to the lack of numerous and recurrent interchanges between video frames. However, the content of measurement M2 was a high motion animated video stream, which made network impairments – especially packet loss – straightforwardly visible to the evaluator, usually in form of artifacts [10].

Terminals are also important variables of usage scenarios. The work of F. Agboma et al [11] details the correlations between the terminal of a given service and perceived quality. Indeed, while the conventional 2D PC monitor display of M1 posed no issues of technology acceptance, sometimes the active 3D display of M2 caused a real headache, literally [12].

Evaluation itself can be done in a qualitative or a quantitative matter. The paper of P. Brooks et al. [13] marks the importance of quantitative evaluation methodologies, since qualitative labels can question objectivity. A qualitative approach may indeed create distortions in evaluation due to the subjective meanings of different labels [14] and may result contrarily in different languages [15]. The paper of A. Watson et al. [16] also doubts the usability of such scales, indicates limitations and warns about the compression of measurement results to the lower half of the scale. V. Menkovski et al [17] argues with the qualitative absolute scale of ratings as well due to the uttermost subjectivity in their interpretations. Both of our measurements utilize the numerical evaluation of discrete scales.

Another work of V. Menkovski et al [18] also emphasize with the dense variety of factors responsible for the non-linear relationship between the physical and the perception domain. They present an active learning algorithm, an adaptive MLDS (Maximum Likelihood Difference Scaling) to increase the efficiency, scalability and the learning rate of the existing approaches [19]. The numerical results of both our subjective measurements contain psychometric functions, putting in relation physical and psychological scales. However, even with a greater number of participants compared to what we have had during measurement M2 (90 participants), variance and bias are expected to be included [20]. Objective solutions like MLDS are not only interesting due to the reduction of bias, but for the elimination of socio-psychological alterations as well.

In our study we deal with this specific type of assessment distortion. Evaluation during a measurement is nothing but a series of decisions. Due to this fact, cognitive dissonance [6] and especially post-decision dissonance [21] affect evaluator behavior. As mentioned earlier, cognitive dissonance is a disharmonic state between conflicting cognitions, which needs to be resolved in order to avoid discomfort, stress and other unwanted feelings. This is quite relevant in case of quality assessment since it encourages test participants to support prior ideas regarding the service instead of perception, resulting in the alteration of the actual experience and thus the scores as well. Post-decision dissonance protects the validity of prior decisions, which in case of assessment, forges a harmony between the results of evaluation tasks; evaluating a given test case in a measurement series is heavily affected by earlier evaluation decisions.

This interesting topic is investigated by others as well. The work of A. Sackl et al. [22] demonstrates the inevitable role of cognitive dissonance in QoE and underlines the correlation between experienced service quality and pricing. There is indeed a close linkage between quality perception and willingness-to-pay, and with the detailed phenomenon of post-decision dissonance, referred to as “post purchase cognitive dissonance”, they managed to clearly explain the background of their results. They emphasize the human action of justification, which is also a key element of the naissance of measurement result distortion in our study. While their first experiment of streaming video evaluation involved real-world currency and active user decisions, the second one lacked interaction. In order to justify the binary decisions of purchasing or not purchasing in the first experiment, the participants evaluated the given services with higher scores compared to the results of experiment without user decisions. Our works are rather related to the second experiment, since the only so-called interaction is the participation in a video conference in M1, as shall be seen later on. However, we still deal with justification in our

series of measurements, due to the presence preconceptions; once an evaluator supports a specific idea with a finalized decision, it is likely to be repeated later on with the purpose of justifying the previous one.

The publications of M. O'Neill and A. Palmer [23][24] also gained our attention. Their research includes a time difference of one month, which enables post-decision dissonance to have a more significant, evolved impact. The intervals between evaluations in our measurements were merely a couple of minutes, resulting in short-term consequences of the phenomenon.

## III. MEASUREMENT CONFIGURATION

### A. Measurement methodology

As mentioned in the introduction, QoE monitoring plays an essential role in designing, initializing and maintaining services. The standard techniques for such measurements are defined by the recommendation [4] of the International Telecommunication Union. It contains all the important parameters that can be involved in the configuration of a QoE measurement. Subjective determination of transmission quality can be achieved by four different clusters of methods. The most popular ones are considered to be the conversation-opinion tests, since they are designed to replicate actual usage of two-way interactive services. Listening-opinion tests rather focus on ones perception, which makes them excellent to measure basic usability and acceptance. Interview and survey tests are efficient methods to extract information beyond a numerical judgment. A group labeled "other tests" is also defined. We decided to use conversation-opinion tests in measurement M1 and listening-opinion tests in M2, both with minor additions from interview tests methodology; test subjects were able to detail their decisions during recorded interviews. Additional verbal extension of evaluation supports understanding the motivations behind evaluator behavior.

Before the measurement itself, the Level of Comprehension of each subject was revealed by asking a set of questions related to the background of the concerned telecommunication technologies and solutions. For instance, in case of M2, questions on 3D display technologies and network security were necessarily included. These conversations, each taking approximately thirty minutes, were recorded for further analysis to precisely determine the LoC of the subjects. It needs to be noted that although this method of LoC determination required vast resources, we could not risk losing a desired level of accuracy. The determination process happened manually in an iterative manner; the ones with the greatest and the lowest technical competences were selected and the process was repeated until all participants were categorized. Although this method prevents the possible LoC overestimation, we shall use a more cost effective approach in the future.

Three different levels were distinguished; level –1 represented the group of those with the lowest, while level +1 represented the highest level of technical comprehension, and level 0 was in between. For more intense investigation of the correlation, more levels could be defined (e.g., M. R. Quintero et al. defined 6 [25]). To preserve the purity of LoC determination, the subjects were given no information about the nature of the measurement before it had begun. The variety of technical competence was not the only aspect during the selection of the test subjects, but it was also necessary to only select people who have never seen each other before in order to prevent information leak between measurements. The subjects haven't even met each other during the series of measurements, because of the different dates and times of the measurements. If any subject had received even the slightest information about the measurement before its date, it could have and probably would have resulted in LoC overestimation.

### B. Configuration of Measurement M1

The basic set of measurements for our analysis of M1 was built on a video conference between the test moderator and the test participant, such emulating a typical mobile video service. The tests were performed on the laboratory network (see *Figure 1*) of the Mobile Innovation Centre [26]. Twenty test subjects participated in the series of measurements with different levels of prior technical knowledge, ranging from simple inexperienced user to IT engineer with PhD degree. Although test subject number may be considered to be low in the aspect of representative results, it is sufficient at this initial phase to expound the phenomenon and analyze evaluator behavior.

The complete process of a measurement was divided into four sections, following each other without delay. The first part was the LoC level determination conversation, as mentioned before. This was extended by questions on general user behavior, involving the quality of previously experienced video conferences. After the basic instructions, began the third and most important part of the process, the mobile video conference and its evaluation. This was concluded by an oral evaluation of the experienced quality, which was also recorded like the first two conversations. The test moderator was the same in each and every part of the process and for all subjects.

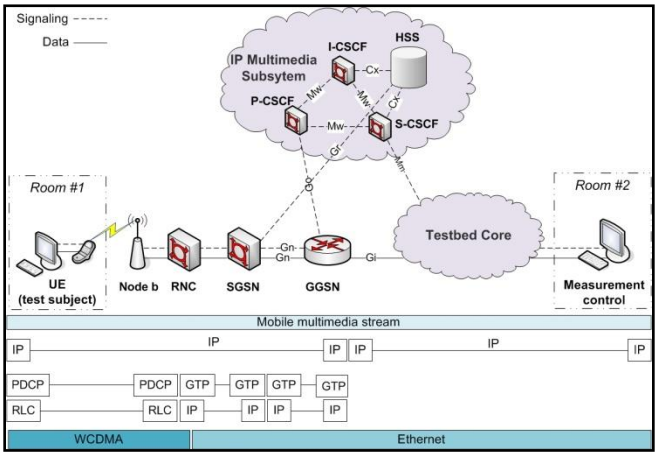


Figure 1. Network topology of measurement M1

During the video conference, the test moderator used a terminal in the laboratory of the Mobile Innovation Centre (Room #2 in *Figure 1*), while the subject was isolated in the conference room of the laboratory. The audiovisual connection was established by a *Linphone 3.2.1* client [27] on an *Ubuntu 10.04* operation system. Both end terminals shared the same hardware and software, including multimedia equipment such as web camera and headset. Connection to the test network, however, was different. While the terminal at the laboratory connected via Ethernet, the computer at the conference room (Room #1 in *Figure 1*) used a Huawei 3G HSDPA wireless modem. IP Multimedia Subsystem (IMS) [28] was in control of the mobile multimedia traffic over the UMTS network.

The complete video conversation took approximately one hour. Although it was divided into twenty subsections (referred to as test cases), the conversation itself was fluent and natural. It was enough to have test cases with 3 minutes of length, since longer test cases would not have led to significant differences in the perception of quality [29]. However, perception varies over time [30], so it was necessary to keep the complete length at a reasonable extent in order to comply with the attention span. Every subsection had a different artificial one-way QoS parameter load in terms of delay, jitter and packet loss, in addition to the real QoS values of the network. To achieve this, we used the command line based *netem* application [31] in order to change the output traffic of the laboratory terminal without the interruption or pause of the video conversation. The achieved impairment of QoS resulted different artifacts and stalling. The parameter values were given to the subject before commencing the conversation, in a form of a QoS parameter matrix (see *TABLE I*), together with the fix parameters of the measurement (see *TABLE II*), such as video resolution. The objective of the subject was to separately evaluate the audio and video quality of the twenty different test cases on a scale from one to ten, where ten represented the best score. Although five-point scales are indeed more popular in case of evaluation, we chose this size in order to support test subjects in distinguishing their experiences.

TABLE I. QOS PARAMETER MATRIX VARIABLE VALUES OF M1

| **Test case** | **Varying parameters** | | |
|---|---|---|---|
| | ***Additional delay*** | ***Additional jitter*** | ***Additional packet loss*** |
| 1 | 0 ms | 0 ms | 0 % |
| 2 | 50 ms | 10 ms | 0.5 % |
| 3 | 200 ms | 40 ms | 2 % |
| 4 | 800 ms | 180 ms | 8 % |
| 5 | 0 ms | 180 ms | 8 % |
| 6 | 0 ms | 0 ms | 8 % |
| 7 | 0 ms | 180 ms | 0 % |
| 8 | 800 ms | 0 ms | 0 % |
| 9 | 800 ms | 100 ms | 1.2 % |
| 10 | 400 ms | 100 ms | 1.2 % |
| 11 | 200 ms | 100 ms | 1.2 % |
| 12 | 100 ms | 100 ms | 1.2 % |
| 13 | 100 ms | 180 ms | 0.5 % |
| 14 | 100 ms | 100 ms | 0.5 % |
| 15 | 100 ms | 40 ms | 0.5 % |
| 16 | 100 ms | 20 ms | 0.5 % |
| 17 | 200 ms | 20 ms | 0.5 % |
| 18 | 200 ms | 20 ms | 2 % |
| 19 | 200 ms | 20 ms | 4 % |
| 20 | 200 ms | 20 ms | 8 % |

TABLE II. DESCRIPTORS OF VIDEO CONFERENCE IN M1

| |
|---|
| Resolution: 640x480 |
| Video codec: MPEG4 |
| Audio codec: speex |

## C. *Configuration of Measurement M2*

QoE measurement series M2 was the assessment of 3D multimedia streams on a PC with Nvidia Vision active 3D technology [32]. The task of the participants was to rate five different aspects of quality of 20 test cases (see *TABLE III*). The chosen aspects were video continuity, image quality, 3D experience, audio/video synchronization and the overall experience. The variables of the test cases included jitter, packet loss, transmission power, and the binary presence of bandwidth limitation and network security.

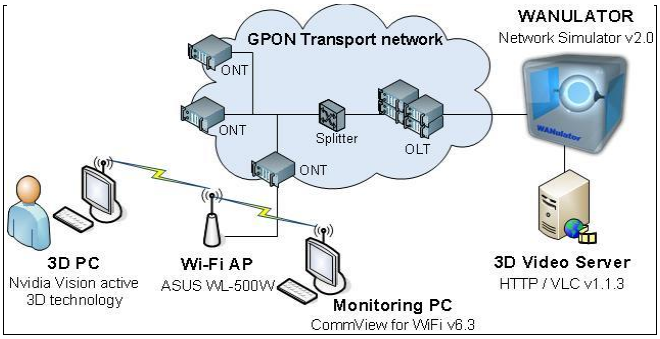


Figure 2. Network topology of measurement M2

The one-minute-long multimedia contents of measurement M2 (see *TABLE IV*) were streamed from a video server and delivered through a GPON network [33], which was accessed from client side via Wi-Fi (see *Figure 2*). While the simulation of varying network parameters was

performed by WANulator on a separate computer, the intensity of transmission power was adjusted on the Wi-Fi AP.

TABLE III. QOS PARAMETER MATRIX VARIABLE VALUES OF M2

| Test case | Varying parameters | | | | |
|---|---|---|---|---|---|
| | *Security* | *TX Power* | *Jitter* | *Packet loss* | *Bandwidth limitation* |
| 1 | NO | 71 mW | 30 ms | 0 % | NO |
| 2 | NO | 71 mW | 0 ms | 1 % | NO |
| 3 | NO | 71 mW | 60 ms | 1 % | NO |
| 4 | NO | 71 mW | 30 ms | 1 % | NO |
| 5 | NO | 71 mW | 60 ms | 0 % | NO |
| 6 | NO | 71 mW | 30 ms | 0 % | YES |
| 7 | NO | 71 mW | 60 ms | 2 % | NO |
| 8 | NO | 71 mW | 0 ms | 2 % | NO |
| 9 | NO | 71 mW | 0 ms | 0 % | NO |
| 10 | NO | 71 mW | 30 ms | 2 % | NO |
| 11 | NO | 71 mW | 60 ms | 0 % | YES |
| 12 | NO | 71 mW | 0 ms | 0 % | YES |
| 13 | YES | 71 mW | 0 ms | 0 % | NO |
| 14 | YES | 71 mW | 30 ms | 1 % | NO |
| 15 | YES | 71 mW | 60 ms | 2 % | NO |
| 16 | YES | 71 mW | 0 ms | 0 % | YES |
| 17 | NO | 35 mW | 0 ms | 0 % | NO |
| 18 | NO | 35 mW | 30 ms | 1 % | NO |
| 19 | NO | 251 mW | 0 ms | 0 % | NO |
| 20 | NO | 251 mW | 30 ms | 1 % | NO |

TABLE IV. DESCRIPTORS OF MULTIMEDIA CONTENT IN M2

| |
|---|
| Resolution: 3360x1050 |
| Video codec: MPEG4 |
| Audio codec: MP3 |

A total of 90 test subjects participated in M2. Similarly to M1, LoC was measured, but only in case of 34 participants. The rest performed so-called blind tests; they did not possess any direct information regarding the differentiation of test cases. The scale of evaluation was a 10-point quantitative discrete scale in this case as well, however, the highest score on the scale carried a slightly different interpretation. While in M1 score 10 was defined as the highest value that can be used for the evaluation of perceived quality, in case of M2 it represented the quality of the reference test case. Although the first test case of M1 can be deemed to be a reference of assessment, since participants were not informed explicitly about its nature, it cannot be considered to be a full-reference subjective QoE measurement, unlike M2. However, M2 also included the reference quality as a subject of evaluation, namely test case 9.

## IV. MEASUREMENT RESULTS

### A. *Results of measurement M1*

Before taking the different LoC levels into consideration, we took a look at the MOS results of M1 (see *Figure 3*). The first thing that grabbed our attention was that test case 8 with its additional 800 ms delay managed to achieve better video scores than the reference test case.

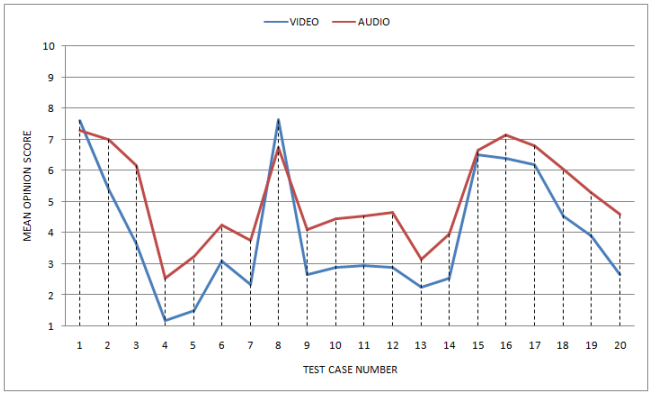


Figure 3. Mean Opinion Score of M1

Although this simulated network impairment rather had its effect on audio quality, it was still perceptible in video quality as well; however, the difference was barely noticeable. How was it possible that a test case with a minor degradation in quality received a higher score than the reference test case? By relying only on the MOS results, it would be quite exigent to give an accurate explanation to this phenomenon. After performing the LoC separation (7 participants in level +1 and –1, 6 participants in level 0) of the results and viewing the recorded video footages, the answer became clear (see *Figure 4*).

While the test participants of LoC level +1 and 0 were commonly controlled by the fact that delay is noxious to experienced quality and thus such measurement case cannot achieve a better score, members of level –1 were not aware of this. In fact, as heard on the recorded conversations, some of them were quite convinced that delay is beneficial and produces a higher level of quality. The other subjects were not affected by such misbelieves so not even a single participant gave test case 8 a better score. The devoted opinion of the evaluators in level –1 on the quality of these two cases was quite sufficient to create a distortion large enough to significantly alter the overall MOS results. It also needs to be noted that participants of level +1 indicated the difference in quality with more caution, even though their preconceptions were more reinforced by their technical knowledge and experience; many of them were more confident that they managed to detect the barely noticeable dissimilarities between the test cases, but they only distinguished them by a single unit on the measurement scale.

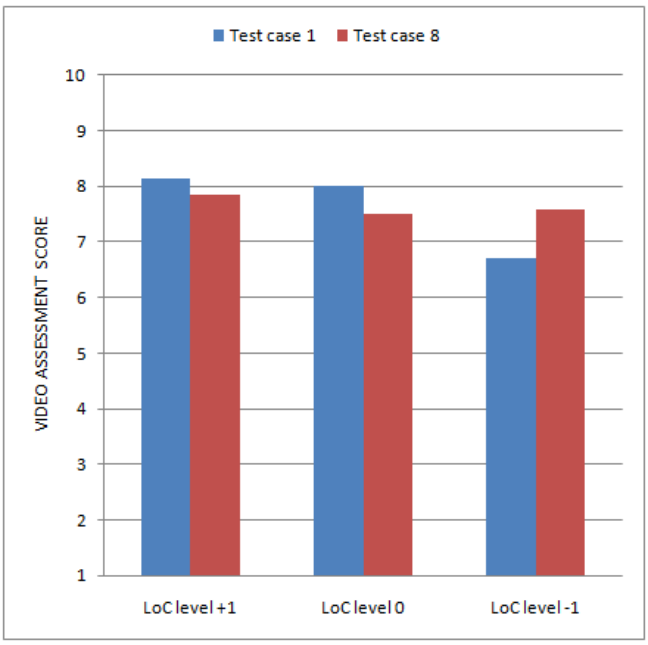


Figure 4. Video assessment scores of test case 1 and 8

Let us approach this issue from the angle of cognitive dissonance. On one hand, perception of many participants were not able to identify evident distinction between the two test cases, while on the other hand, preconception contained a clear direction of the difference. This dissonant state of cognitions was solved by either the alteration of perception ("I can clearly see the difference") or the reconsideration of its correctness ("I cannot clearly see the difference but I know it has to be there"). Those in LoC level –1 who supported the preconception of a beneficial delay in the aspect of video quality were fuelled by post-decision dissonance during the evaluation of test case 9 to 12. In these four test cases delay was reduced while the other parameters remained the same. Again, there were only the slightest differences in video quality, yet they made a decreasing score pattern, since preconception was also aided by a prior decision.

Those who utilized test case 1 as reference quality were strictly bounded by the rule that no other test case could ever exceed its score. However, only two participants from LoC level +1 granted it the maximal 10 points. This is a natural behavior when using an evaluation scale. Participants did not wish to limit the expressive ability of their evaluations; by using the top or the bottom end of the scale – especially during an early test case – participants forfeit the chance to express their thoughts should a test case with even greater or lesser quality appear. However, this implies the sacrifice of evaluation space; such participants were limited to use a 9-point or smaller scale. This idea also resulted these participants forcefully gave lower scores to each and every test case, even when no evident difference was found, as detailed earlier.

On the other hand, those who were not bounded by this test case were able to rate other test cases higher than the first one. The series from test case 13 to 16 was the reduction of the amount of additional jitter (see *Figure 5*). There was an immense difference between test case 14 and 15; while the video image of test case 14 was barely recognizable, test case 15 provided an acceptable video quality. This dire change of quality motivated some participants of LoC level –1 to give high scores, higher than test case 1.

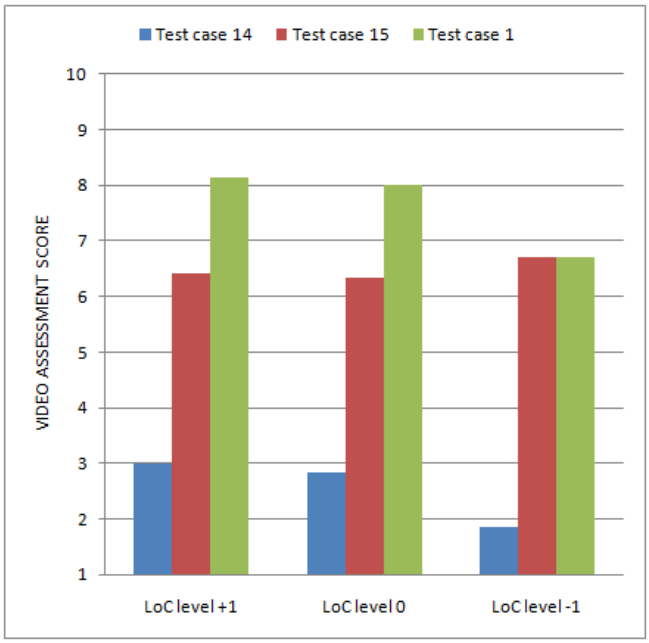


Figure 5. Video assessment scores of test case 14 and 15

The first four test cases represented a general decrement in QoS values; both delay, jitter and packet loss were increasing. In this case, it was quite interesting to see that the higher LoC level a participant had, the closer his/her evaluation was to uniformity (e.g., 10, 7, 4, 1) in both video and audio quality.

In audio quality evaluation scores, the progress from test case 9 to 12 was possibly the most interesting. These four test cases endured delay reduction while preserving a notable constant jitter. Presuming the experienced quality tendencies during these four test cases is not a trivial task. It was beneficial to have a smaller delay, however, the ratio of jitter and delay increased. The audio MOS shows a definite raise, even though none of the participants thought it that way. In level –1 and 0, there was no repeating behavior pattern. In fact, participants used a high variety of scoring patterns to assess, since there was no obvious difference in the overall experience of audio quality. On one hand, mutual speech interruptions were fewer, but on the other, audio quality was less enjoyable to some extent. The scores given by the participants were based on the personal decision whether the first or the second effect was more dominant. However, the audio assessments in LoC level +1 were shocking; 6 out of 7 participants used a constant evaluation pattern (see *Figure 6*). It means that preconceptions had such a high level impact on evaluation that these subjects ignored any lesser differences that they experienced between cases. They

considered the opposing effects nearly equal, which supposes an unvarying overall experience.

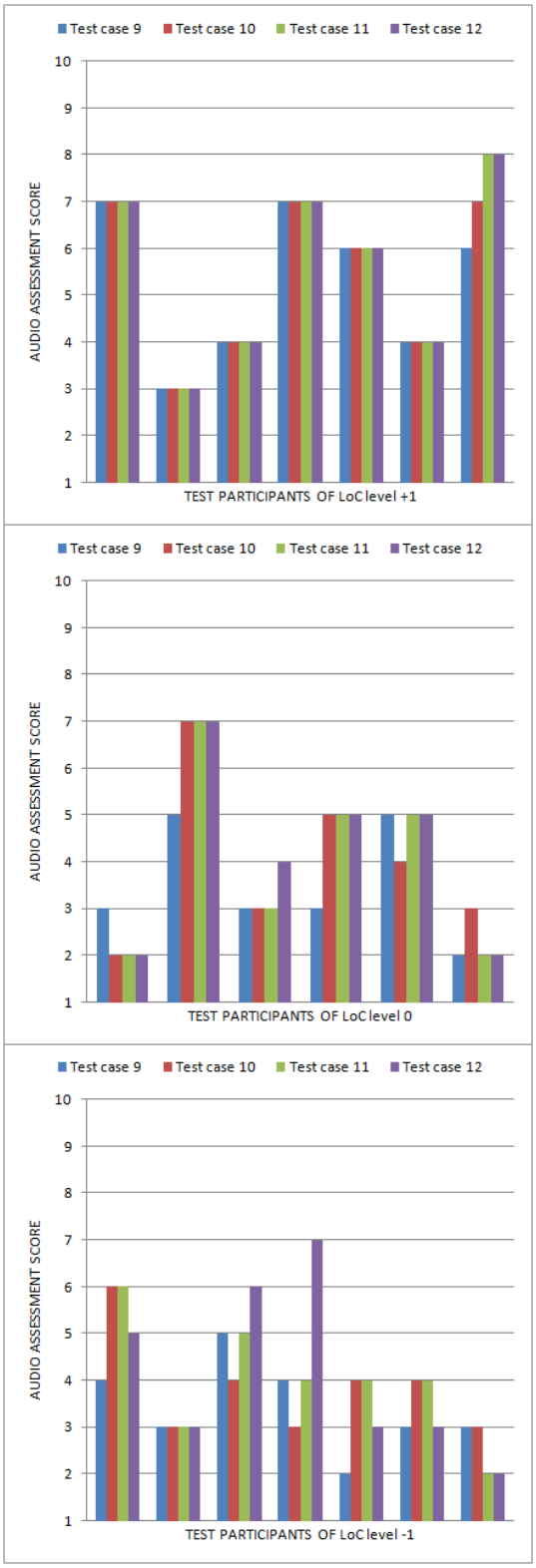


Figure 6. Audio assessment scores of test cases 9 to 12

A participant from LoC level –1 also used constant evaluation, but claimed that "jitter has no effect on audio quality".

It is also exciting to compare the audio results of test case 17 and 18. Due to the redundancy of the human voice, the given amount of packet loss caused no major difference between these two test cases. Many members of LoC level +1 and some of 0 indicated an apparent difference in scoring, since according to their preconceptions, audio quality should clearly lessen. However, there were participants in level –1 with the idea that packet loss is beneficial in the aspect of audio quality. This is also a great example for the disobedience of subjective prior cognition, since their scoring direction was inverted in the last two test cases. Even though preconception was supported by post-decision dissonance through a prior decision, the test participants had to abandon it when facing the obviously lessening sound quality of test case 19 and 20. Their assessment was so intense that it managed to make a clear impact on the audio MOS.

### *B. Results of measurement M2*

This subsection mainly focuses on the assessment of the 34 participants with access to the alteration of the variables. The psychometric functions of the other results [34] are indeed also exciting, but this paper emphasizes more with the effects of direct environmental information. Moreover, this paper does not deal with the separated aspects of quality, but uses weighted averages, since participants were asked to weight these aspects based on personal importance with the sum of 10 (2 for each if all are equally important).

We approached the results of M2 from five directions: security presence, transmission power adjustment, jitter, packet loss and the limitation of bandwidth (see *TABLE V*). If we take a look at the MOS (see *Figure 7*), the tendencies of the evaluation results of the groups of participants with and without access to environmental information might seem to differ at some points.

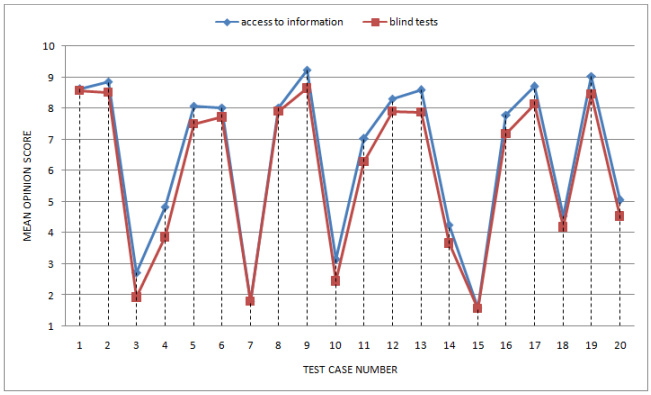


Figure 7. Mean Opinion Score of M2

If we just approach these data without any of the previously mentioned directions, the first thing we notice is that the mean assessment of those with QoS awareness is higher in scores. For instance, test case 9 – the hidden

reference test case – achieved better evaluation results due to the fact that participants were aware that it was without any additional load.

TABLE V. QOE RESULTS OF THE INVESTIGATED ASPECTS IN M2

| *Test case* | *Investigated aspect* | *MOS (blind tests)* | *MOS (with access)* |
|---|---|---|---|
| 19 | Transmission power | 8.46 | 9.04 |
| 17 | | 8.13 | 8.7 |
| 20 | Transmission power | 4.53 | 5.05 |
| 18 | | 4.19 | 4.54 |
| 1 | Bandwidth limitation | 8.57 | 8.61 |
| 6 | | 7.72 | 8.02 |
| 5 | Bandwidth limitation | 7.49 | 8.08 |
| 11 | | 6.29 | 7.02 |
| 9 | Bandwidth limitation | 8.64 | 9.24 |
| 12 | | 7.91 | 8.3 |
| 13 | Bandwidth limitation | 7.87 | 8.58 |
| 16 | | 7.19 | 7.79 |
| 9 | Security presence | 8.64 | 9.24 |
| 13 | | 7.87 | 8.58 |
| 4 | Security presence | 3.87 | 4.84 |
| 14 | | 3.66 | 4.26 |
| 7 | Security presence | 1.79 | 1.8 |
| 15 | | 1.57 | 1.6 |
| 12 | Security presence | 7.91 | 8.3 |
| 16 | | 7.19 | 7.79 |
| 1 | Jitter | 8.57 | 8.61 |
| 5 | | 7.49 | 8.08 |
| 6 | Jitter | 7.72 | 8.02 |
| 11 | | 6.29 | 7.02 |
| 4 | Jitter | 3.87 | 4.84 |
| 3 | | 1.93 | 2.71 |
| 10 | Jitter | 2.45 | 3.14 |
| 7 | | 1.79 | 1.8 |
| 2 | Packet loss | 8.51 | 8.86 |
| 8 | | 7.91 | 8.02 |
| 3 | Packet loss | 1.93 | 2.71 |
| 7 | | 1.79 | 1.8 |
| 4 | Packet loss | 3.87 | 4.84 |
| 10 | | 2.45 | 3.14 |

The scores of test case 12 and 13 are also quite interesting; they both had a standard transmission power of 71 mW and no additional jitter or packet loss, but while test case 12 had limited bandwidth, test case 13 utilized secure transmission. The very similar situation can be witnessed in the relationship of test case 5 and 6. The parameters of bandwidth limitation on WANulator were chosen to imply a barely noticeable difference in quality. However, the parameter matrix only included this in a binary way, without any exact value. Preconceptions regarding bandwidth limitation were quite amplified, usually regardless of LoC level, since the majority depicted the word "limitation" as something harmful to quality, which it actually is.

In M2, the adjustment of additional jitter and packet loss had a rather evident effect on the experienced quality of the 3D stream transmission. Thus any prior idea of beneficial jitter or packet loss was nullified by perception.

Preconceptions regarding transmission power were a bit more diverse. There were quite some participants who approached the alteration of transmission power somewhat similar to sound volume, where too high is just as adverse as too low.

The information regarding these previous aspects was commonly used in the same way during assessment, apart from a few participants, whose evaluation scores did not affect the mean QoE results of their LoC levels. Similarly to M1, 3 different levels of LoC were distinguished (11 participants in level +1 and –1, 12 participants in level 0). The most interesting results appeared when we viewed the security presence aspect scores of LoC separation.

Although the mean results of those with access to the QoS parameters were similar to the others in scoring relations, it was revealed that there were many participants in LoC level –1 and some in level 0 with a steady preconception stating that secure transmission has to have better performance. It was so influential that it managed to become visible in the mean scores of level –1 (see *Figure 9, 10 and 11*).

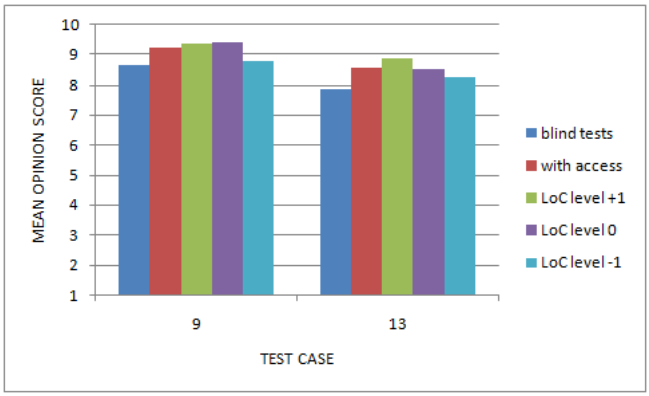


Figure 8. Mean Opinion Score of test case 9 and 13

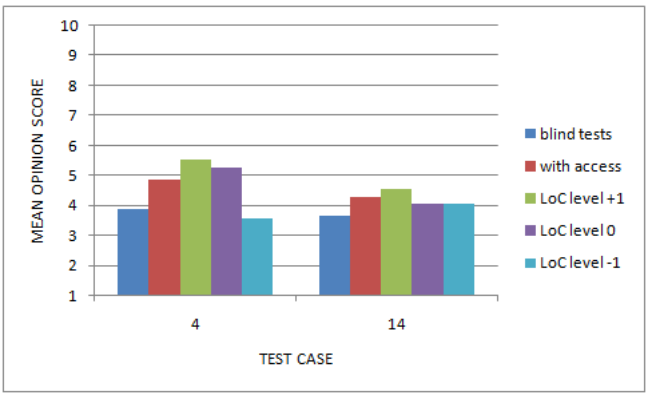


Figure 9. Mean Opinion Score of test case 4 and 14

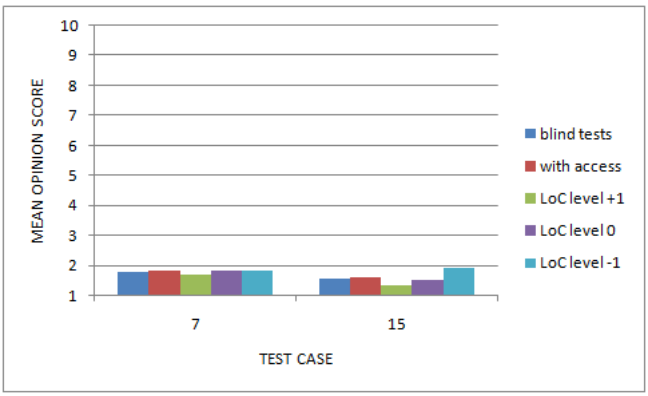


Figure 10. Mean Opinion Score of test case 7 and 15

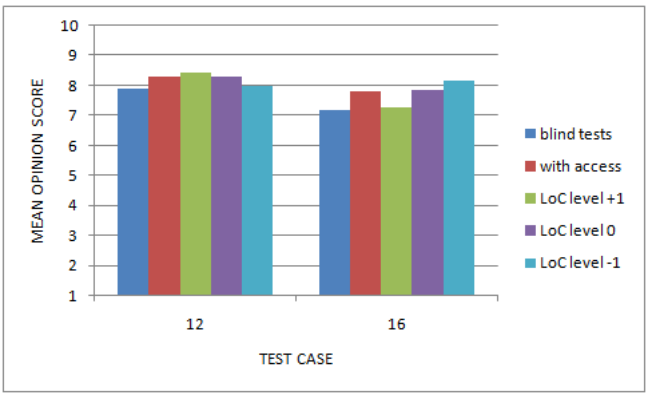


Figure 11. Mean Opinion Score of test case 12 and 16

Even though this phenomenon did not appear between the mean scores of test case 9 and 13 (see *Figure 8*), 3 out of 11 participants already supported that preconception; this number in case of test case 4 and 14 (see *Figure 9*) was 7 out of 11. Of course on the other hand, several participants belonging to LoC level +1 were confident that test cases with secure transmission had to be worse due to technical reasons. The difference of the experienced quality in practice was almost ignorable, which enabled preconception to dominate perception through cognitive dissonance.

## V. Conclusion

The paper presented correlations between assessment alteration and the Level of Comprehension of test participants and detailed the socio-psychological background of the phenomenon. Environmental information regarding the given service can be considered the actual hotbed of preconceptions. Its relevancy is supported by the single fact that the majority of evaluation measurements cannot be considered to be so-called blind tests due to their configurations. The presented measurement utilized a radical amount and type of information, usually not public during service assessment and everyday service usage. However, in many cases basic information – like the type of connection – is very hard or impossible to hide.

Currently our researches deal with hard-to-hide environmental information, which are naturally present to evaluators. In the upcoming measurements, the methods for LoC determination will be simplified, however, at this initial phase of the research series we couldn't risk to lose any level of accuracy. Our future goals also contain the exhaustive analysis and comparison of automated and human assessment of quality, since objective solutions are invulnerable to the distortions presented in this paper.

## Acknowledgement

The research leading to these results has received funding from the European Union's Seventh Framework Programme ([FP7/2007-2013]) under grant agreement n° 288502. This work was also supported by the Mobile Innovation Centre Hungary (MIK). We are grateful to the Department of Networked Systems and Services (HIT) and to the Department of Telecommunications and Media Informatics (TMIT) of the Budapest University of Technology and Economics (BME). We would also like to thank Ivett Kulik for her help and cooperation. Last but not least we would like to thank the many individuals whose work made this research possible.

## References

[1] Qualinet White Paper on Definitions of Quality of Experience. March 2013. http://www.qualinet.eu/images/stories/QoE_whitepaper_v1.2.pdf (retrieved January 2014)

[2] ITU-T Rec. E.800, Definitions of terms related to quality of service. Int. Telecomm. Union, Geneva, September 2008.

[3] M. Fiedler, T. Hossfeld, P. Tran-Gia. A generic quantitative relationship between quality of experience and quality of service. Network, IEEE , vol.24, no.2, pp.36–41, March-April, 2010.

[4] International Telecommunication Union. Methods for subjective determination of transmission quality. ITU Recommendation P.800 (08/96), August 1996.

[5] P. A. Kara, L. Bokor, S. Imre. Distortions in QoE measurements of ubiquitous mobile video services caused by the preconceptions of test subjects. IEEE/IPSJ International Symposium on Applications and the Internet SAINT2012. Izmir, Turkey, July 2012. pp. 409–413.

[6] L. Festinger. A theory of cognitive dissonance. Stanford, CA: Stanford University Press, 1957.

[7] International Telecommunication Union. P series: Terminals and subjective and objective assessment methods. ITU-T Recommendations, P series. http://www.itu.int/rec/T-REC-P/en (retrieved January 2014).

[8] I. Ketykó, K. De Moor, W. Joseph, L. Martens, and L. De Marez. Performing QoE-measurements in an actual 3G network. In IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB 10), pp. 1-6, March 2010.

[9] G. Exarchakos, L. Druda, V. Menkovski, P. Bellavista, A. Liotta. Skype resilience to high motion videos. International Journal of Wavelets, Multiresolution and Information Processing, Vol.11(3), 2013, World Scientific Publishing.

[10] C. T. E. R. Hewage, M. G. Martini. Quality of experience for 3D video streaming. Communications Magazine, Volume 51, Issue 5, pp.101–107, May 2013.

[11] F. Agboma, A. Liotta. Quality of Experience Management in Mobile Content Delivery Systems, Journal of Telecommunication Systems, special issue on the Quality of Experience issues in Multimedia Provision. Vol. 49(1), pp. 85-98, Springer 2012.

[12] I. Kulik, P.A. Kara, T.A. Trinh, L. Bokor. Analysis of the Relationship between Quality of Experience and Service Attributes for 3D Future Internet Multimedia. IEEE 4th International Conference on Cognitive Infocommunications, Budapest, Hungary, 2-5 Dec. 2013, pp. 641–646.

[13] P. Brooks, B. Hestnes. User Measures of Quality of Experience: Why Being Objective and Quantitative Is Important. Network, IEEE, Vol. 24, No. 2. (March 2010), pp. 8–13.

[14] Jones, BL. & McManus, P.R. Graphic scaling of qualitative terms. SMPTE Journal, November 1986, pp. 1166–1171.

[15] Narita, N. Graphic scaling and validity of Japanese descriptive terms used in subjective-evaluation tests. SMPTE Journal, July 1993, pp. 616–622.

[16] A. Watson, M. A. Sasse. Measuring Perceived Quality of Speech and Video in Multimedia Conferencing Applications. In MULTIMEDIA '98: Proceedings of the sixth ACM international conference on Multimedia (September 1998), pp. 55–60.

[17] V. Menkovski, G. Exarchakos, A. Liotta. The Value of Relative Quality in Video Delivery. Journal of Mobile Multimedia, Vol.7(3), pp. 151-162. Rinton Press, September 2011.

[18] V. Menkovski, A. Liotta. Adaptive Psychometric Scaling for Video Quality Assessment. Journal of Signal Processing: Image Communication. Vol.26(8), pp.788–799. Elsevier. 2012.

[19] C. Charrier, L. T. Maloney, H. Cherifi ,K. Knoblauch. Maximum likelihood difference scaling of image quality in compression-degraded images. Journal of the Optical Society of America A24 (11), 2007, pp. 3418–3426.

[20] A. B. Watson. Proposal: measurement of a JND scale for video quality. IEEEG-2.1.6 Subcommittee on Video Compression Measurements, 2000.

[21] Knox, R. E., & Inkster, J. A. Postdecision dissonance at post time. Journal of Personality and Social Psychology, 1968, pp. 319–323.

[22] A. Sackl, P. Zwickl, S. Egger, P. Reichl. The role of cognitive dissonance for QoE evaluation of multimedia services. 2012 IEEE Globecom Workshops (GC Wkshps), pp. 1352–1356.

[23] M. O'Neill, A. Palmer. Cognitive dissonance and the stability of service quality perceptions. Journal of Services Marketing, 2004, Volume 18, Issue 6, pp. 433-449.

[24] M. O'Neill, A. Palmer. Exploring the relationship between post-consumption dissonance and time-elapsed perceptions of service quality. Anzmac conference, New Zealand, 2001.

[25] M. R. Quintero, A. Raake. Is taking into account the subjects degree of knowledge and expertise enough when rating quality? QoMEX 2012: 194-199.

[26] BME-MIK. Budapest University of Technology and Economics - Mobile Innovation Centre, Official Website. https://www.mik.bme.hu/home/aboutus/, (retrieved January 2014).

[27] Linphone. Official Website. http://www.linphone.org/, (retrieved 2012 May).

[28] 3GPP TS 23.228. IP Multimedia Subsystem (IMS); Stage 2. Rel-8, 2008.

[29] P. Froehlich, S. Egger, M. Schatz, R. Muehlegger, K. Masuch, and B. Gardlo, "QoE in 10 Seconds: Are Short Video Clip Lengths Sufficient for Quality of Experience Assessment?" in Proceedings of the fourth International Workshop on Quality of Multimedia Experience QoMEX, 2012.

[30] E. B. Goldstein. Sensation and Perception, Eighth Edition. Cengage Learning, February 2009.

[31] netem. Linux Network Emulation Official Website. http://www.linuxfoundation.org/collaborate/workgroups/networking/netem, (retrieved January 2014).

[32] NvidiaVision Player website http://www.nvidia.com/object/3d-vision-video-player-1.7.5-driver.html (retrieved January 2014).

[33] I. Kulik, T.A. Trinh. Investigation of Quality of Experience for 3D Streams in GPON. Ralf Lehnert (Ed.) EUNICE 2011. LNCS, vol. 6955, pp.157 – 168 Springer, Heidelberg, 2011.

[34] I. Kulik, P.A. Kara, T.A. Trinh, L. Bokor. Attributes unmasked: Investigation of service aspects in subjective evaluation of wireless 3D multimedia. IEEE/ICIA Second International Conference of Informatics, Lodz, Poland, 23-25 Sept. 2013, IEEE, pp.270–275.