University of Würzburg
Institute of Computer Science
Research Report Series

Crowdsourced Subjective User Study Results on QoE Influence Factors of HTTP Adaptive Streaming

Tobias Hoßfeld, Michael Seufert,
Thomas Zinner[1] and Christian Sieber[2]

Report No. 491 July 2014

[1] University of Würzburg
Institute of Computer Science
Chair of Communication Networks
Am Hubland, 97074 Würzburg, Germany
tobias.hossfeld@uni-wuerzburg.de

[2] Technische Universität München
Institute for Communication Networks
Arcisstrasse 21
80290 München, Germany
c.sieber@tum.de

NOTE: The technical report is an extended version of the QoMEX 2014 paper containing relevant detailed information on the technical setup of the measurement studies and the test conditions. Please cite the peer reviewed paper as follows: *Tobias Hoßfeld, Michael Seufert, Christian Sieber, and Thomas Zinner. "Assessing Effect Sizes of Influence Factors Towards a QoE Model for HTTP Adaptive Streaming". QoMEX 2014, Singapore 18-20 September, 2014.*

Crowdsourced Subjective User Study Results on QoE Influence Factors of HTTP Adaptive Streaming

Tobias Hoßfeld, Michael Seufert, Thomas Zinner
University of Würzburg
Institute of Computer Science
Chair of Communication Networks
Am Hubland, 97074 Würzburg, Germany
tobias.hossfeld@uni-wuerzburg.de

Christian Sieber
Technische Universität München
Institute for Communication Networks
Arcisstrasse 21
80290 München, Germany
c.sieber@tum.de

Abstract

HTTP Adaptive Streaming (HAS) is employed by more and more video streaming services in the Internet. It allows to adapt the downloaded video quality to the current network conditions, and thus, avoids stalling (i.e., playback interruptions) to the greatest possible extend. The adaptation of video streams is done by switching between different quality representation levels, which influences the user perceived quality of the video stream. In this work, the influence of several adaptation parameters, namely, switch amplitude (i.e., quality level difference), switching frequency, and recency effects, on Quality of Experience (QoE) is investigated. Therefore, crowdsourcing experiments were conducted in order to collect subjective ratings for different adaptation-related test conditions. The results of these subjective studies indicate the influence of the adaptation parameters, and based on these findings a simplified QoE model for HAS is presented, which only relies on the switch amplitude and the playback time of each layer.

1 Introduction

Video streaming has evolved to the dominating application in the current Internet, and its share is expected to grow even further within the near future [1]. Over-the-top (OTT) video distribution networks like YouTube, Hulu, or Netflix typically use a HTTP/TCP progressive streaming approach. This allows for the use of the advantages of HTTP, i.e., the HTTP delivery structure, an easy network address translation (NAT) and firewall traversal, as well as the advantages of TCP, i.e., congestion control and guaranteed packet delivery. The buffering of content at the client's end further allows to overcome limitations of network resources on short time scales and to assure a continuous play out of the video content. If this is not possible, e.g., in case of live video streaming, limited network resources may lead to buffer underruns and the interruption of the playback.

To overcome this problem and to allow for a flexible adaptation of the video quality to the available network resources and device capabilities, HTTP Adaptive Streaming (HAS) has been designed. The video content is available in multiple bit rates, i.e., quality levels, and split into small segments each containing a few seconds of playtime. The client measures the current bandwidth and/or buffer status and requests the next part of the

video in an appropriate bit rate such that stalling is avoided and the available bandwidth is best possibly utilized. Hence, the control intelligence which segment to stream has moved from the servers to the clients. The HAS streaming technology is adopted by a wide range of applications and video content providers [2] and is standardized in ISO/IEC 23009-1 [3].

Much research in the HAS area tries to find the best downloading strategy in order to maximize the user perceived quality. Influence parameters which are typically investigated are the initial delay, stalling delays and frequencies, the played back video quality as well as the time on a video quality and the switching frequency. Based on the current network conditions, the video characteristics, different video codecs and the monitoring parameters taken into account, the HAS adaptation algorithm tries to maximize the user's QoE. Several download algorithms for HAS have been proposed recently, both, for single layer and multi layer codecs. These algorithms either tend to be optimistic, i.e., they quickly switch to the best possible video quality and accept oscillations, or to be conservative, i.e., they stick to a low video quality and avoid oscillations. The impact of these effects on the user-perceived quality is not fully understood yet, and a user-centric comparison of the different algorithms is missing.

The contribution of this paper is a subjective investigation of the influence of several adaptation parameters, namely, switch amplitude (i.e., quality level difference), switching frequency, and recency effects on Quality of Experience (QoE). Therefore, we conducted crowdsourcing experiments to collect subjective ratings for different adaptation-related test conditions. The results of the conducted studies allow a quantification of the adaptation parameters. Based on the findings we present a simplified QoE model for HAS, which relies on the switch amplitude and the playback time of each layer.

The remainder of this paper is structured as follows. Section 2 discusses the dimensions of QoE for HAS and presents related work. Detailed information on the experimental setup as well as on the conducted user surveys are highlighted in Section 3. Section 4 presents the results and a simplified QoE model is derived. The paper is concluded in Section 5.

2 Dimensions of HTTP Adaptive Streaming QoE

Most frameworks and definitions (e.g., [4, 5]) highlight that some dimensions are especially important for QoE in general. These levels, namely context, user, system, and content, will be covered in this work on the HAS QoE and are presented shortly next.

Context Level. The context level considers aspects like the environment where the user is consuming the service, the social and cultural background, or the purpose of using the service like time killing or information retrieval. Therefore, the quality of a video is perceived differently whether a user desperately wants to watch a specific clip or whether he is simply browsing videos [6]. In this study, we use a crowdsourcing setup for the subjective quality test. This means, the participants are non-expert users with diverse demographics and background which conduct the test in their typical environment. As

the test execution is unsupervised, other influence factors (e.g., the reliability of the participants) have to be monitored during the test and evaluated afterwards [7].

User Level. The subjective perception of adaptive video streaming is most dominated by the user himself. The user level includes psychological factors like expectations of the user, memory and recency effects, or the usage history of the application. In this study, we investigate recency effects like the last quality level of a clip and the recency time, i.e., the time since the last quality adaptation.

System Level. The system dimension is an abstraction level for the technical parameters of video streaming. They cover influences of the transmission network, the devices and screens, but also of the implementation of the application itself like video buffering and adaptation strategies. In this study, we abstract network disturbances and application behavior into quality switching patterns, based on real-world measurements from a mobile scenario [8].

Content Level. Although the content level is not explicitly mentioned in [5], several works (e.g., [9,10]) found that the content (especially spatial/temporal information) plays an important role how adaptation is perceived. The content level addresses the video codec, format, resolution, but also duration, contents of the video, type of video and its motion patterns. In this study we use a single content, i.e., a 14 seconds video clip from the open-source short movie "Tears of Steel" and three quality levels. As with all studies, no absolute results can be presented because for different videos there are different content influences. Nevertheless, the focus of this study is the identification of general, qualitative influence factors and their effects on QoE of adaptive video streaming.

Related Work. [11] investigated rapid alternation of high and low quality levels in adaptive video streaming to mobile devices. They found three important effects, namely adaptation frequency, quality level amplitude, and content. In [12], the quantization parameter of H.264 video streams was changed during the playback. The authors found that QoE decreases slowly when the quantization parameter (qp) starts to increase, i.e., the video bit rate decreases and the image quality gets worse. However, when a certain threshold was reached, a small increase of the qp parameter resulted in a strong decrease of the perceived quality. [13] found that the adaptation frequency should be kept as small as possible. If adaptation cannot be prevented, its amplitude should be kept as small as possible. Thus, a stepwise decrease of image quality is preferred to one abrupt drop. Although several works already investigated particular aspects of Quality of Experience of HTTP adaptive streaming, a holistic model is still missing. This work aims at providing the foundations for such a model by investigating the influence of different QoE dimensions: amplitude, recency, adaptation frequency, and time on highest layer.

3 Experimental Setup and Subjective Studies

Crowdsourcing Framework and its Implementation. Previous studies foreshadowed that the playback quality and quality level switching frequency may have a significant impact on the perceived quality of experience (QoE) of the user. In the following, we introduce the methodology of the studies we designed to substantiate and extend the previous findings. The crowdsourcing experiment was designed in cooperation with microworkers.com, a provider of crowdsourcing services with international reach and large user base. Workers and task creators both register with microworkers.com and can afterwards browse the available tasks or create tasks, respectively. For each fulfilled task, the anonymous workers receive a monetary compensation from the task creator. The payment process is handled by the crowdsourcing platform. We utilized the web-based QualityCrowd2 framework [14] for our QoE tests and followed the best practices for QoE tests introduced in [15] to increase the accuracy of the results in the face of remote participants. To obtain the results presented in this paper, we conducted multiple user studies with approximately 710 unique microworkers (based on the microworkers.com account ID).

Each test subject had to complete a short demographic survey before the experiment. We asked two questions. First, *What is your main reason for using the Internet?* and second *Which continent do you live on?*. As answers we offered *Professional (at work)* and *For fun at home* for the former and the 6 continents for the later. The results show that the majority of the users accessed the campaign's web-site from Asia (70%) and from Europe (26%) and that the test subjects primarily access the Internet from work (64% at work, 36% at home).

After the test subject completed the survey, an introduction in simple English and illustrated with pictures was presented to the participant. The introduction explained how to use the provided web-interface to watch and rate the test sequences. Following the introduction, the test sequences were presented to the participant in sequential order. To prevent any abuse (e.g., fast skipping through the sequences) or problems related to insufficient bandwidth (e.g., stallings), the test sequences were first downloaded to the browser cache and the user had to watch and rate the current sequence before being able to go to the next one. After the playback of the video sequence, the user was asked *Did you notice any changes in quality during playback? If yes, did you feel annoyed by them?* and was presented a 5-point ACR slider with the options *Imperceptible (did not notice any)*, *Perceptible but not annoying (did notice, but did not care)*, *Slightly annoying*, *Annoying* and *Very annoying*.

Video Contents and User Rating. As known from literature, content plays a key role in QoE. Additional HAS QoE influence factors considered in related work are often switching amplitude, switching frequency and recency effects, while the time on high layer is often neglected. Therefore, we focus especially on the time on highest layer, while the combined investigation of different contents and time on highest layer remains for future work.

As test sequence a 14 second (336 frames, starting from time-stamp 00:00:25) video from the movie "Tears of Steel", an open-source short movie produced and published by the Blender Foundation, was used. The scene depicts two persons standing on a small bridge and contains a low level of detail and motion (SI: 8.5, TI: 5.37). We encoded the test sequence using H.264/AVC (libx264) with QP=24 and 24 frames per second into three quality levels by downscaling the source material to 640x360, 320x180 and 160x90. The audio channel was copied from the source video. The encoded videos with audio as used in the crowdsourcing user studies had an average bitrate between 0.64 $^{\text{Mbit}}/_{\text{s}}$ and 0.75 $^{\text{Mbit}}/_{\text{s}}$, i.e. about 1.1 MB to 1.3 MB per video. In the browser of the user, the three quality levels were all upscaled and downscaled, respectively, to a size of 640x360.

In particular, the process used to generate the presented sequences can be summarized as follows. First we downloaded the 720p version of the movie from the official homepage and decoded it into uncompressed images with a width of 1280 and a height of 534 pixels (1280x534). Afterwards black bars (each a height of 93 pixels) were added to the top and bottom to increase the image size to 1280x720. In the next step we downscaled the uncompressed images to two different resolutions, 160x90 (i.e. low quality level) and 640x360 (i.e. high quality level). Next, the low quality image sequence was upscaled to the size of the high quality level (i.e. 640x360). Afterwards we used a script to generate the desired switching patterns by choosing for each frame the desired quality level from the two equally sized image sequences. In the final step, all resulting uncompressed image sequences were encoded with equal parameters for the viewing in the Qualitycrowd2 framework. More technical details on the implementation and the video test sequences[1] can be found in [16].

The crowdsourcing experiment was designed to evaluate five possible QoE influence factors of HAS as shown in Table 1. For each effect we created multiple test sequences with different switching patterns (e.g., number of switches: 1, 2, 3, 4, 5, 6, 8 and 14 switches). Each pattern was reliably rated by at least 82 users and on average over all user studies by 106 participants. Each subject rated between 7–9 test sequences, cf. [**?**] for detailed information.

We also created three test sequences (i.e. one for each quality level) with no quality switches and included them in the conducted user studies as reference for the evaluation of the results. The results show an average MOS value of 4.14 (95 % confidence interval 4.09 to 4.18) for the highest layer based on the ratings of the 710 participants in all user studies. For the lowest layer test sequence included in some of the studies we observed an average MOS value of 2.51 (95% confidence interval of 2.37 to 2.66) based on 267 user ratings. The quality layer in between received a MOS rating of 3.52 (95% confidence interval of 3.42 to 3.62) out of 310 user ratings.

Filtering of Data and Unreliable User Ratings. In order to obtain reliable results despite the lack of control over the anonymous workers, methods were designed and deployed to counteract cheating and unreliable test subjects as suggested in [15]. A number of anti-cheating mechanisms were already built into the QualityCrowd2 frame-

[1]The scripts for generating the test video sequences are available at `http://git.io/hfCaZg`.

work. For example, the test subjects have to watch the whole sequence and move the rating slider to be able to proceed to the next test sequence. In addition to the built-in mechanisms, we showed the participants one simple content questions during the experiment. We asked *Where did the protagonists stand on?* and offered *A building*, *A large field*, *A small bridge* and *Riding on an elephant*. Test participants failing that content questions were discarded from the results. Over all user studies, we observed that 11% of the participants failed the simple content question.

4 Numerical Results

Due to the crowdsourcing setting, user studies were split into several smaller crowdsourcing tasks, as the task duration of crowdsourced QoE evaluation should be in the order of minutes [17]. Therefore, the influence factor analysis and the result presentation is decoupled from the crowdsourced user studies. Detailed information on all crowdsourcing campaigns and the test video sequences can be found in the appendix. A summary of the crowdsourcing experiments is given in Table 1.

Effect	Test design
Amplitude	Switch amplitude high or low and different number of switches
Last Quality Level	End on high or low quality level for different number of switches
Recency Time	Different recency times for different number of switches
Frequency	Different number of switches for constant time on high layer
Time on Highest Layer	Different time on highest layer for different number of switches

Table 1: Summary of crowdsourcing experiments for HAS QoE.

4.1 Amplitude Effect

In order to investigate the effect of switching amplitude, i.e., the quality level difference, on adaptive video streaming, two different amplitudes were compared. Five switching patterns (from N=1 to 8 switches) were tested both with low and high amplitude

switches. To be more precise, each video started at the highest quality and at each switching event the quality oscillated between the highest quality and a medium quality level (low amplitude) or a poor quality level (high amplitude). Figure 1 shows the results of this experiment. On the x-axis the number of switches, and on the y-axis the mean opinion scores (MOS) and the 95% confidence intervals are displayed. The dark bars indicate a low amplitude, i.e., an oscillation between highest and medium quality, whereas the light bars represent a high amplitude, i.e., an oscillation between highest and lowest quality. It can be seen that for all switching patterns, a low amplitude clearly outperforms a high amplitude. Low amplitude patterns always reach a 0.5-1 point higher MOS rating than the corresponding high amplitude patterns without overlapping confidence intervals. This means, the effect of amplitude is significant and has to be taken into account for the QoE of adaptive video streaming. In the following, we will use conditions which contain a high amplitude.

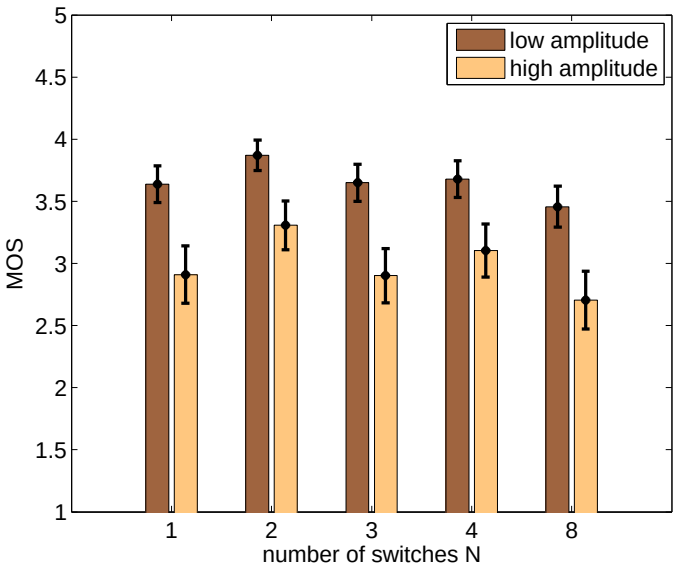


Figure 1: Amplitude Effect is main effect (see Table 2). Two-sample t-test shows significant differences ($p = 0$) between groups (low vs. high amplitude) with an effect size in terms of Cohen's $d = 0.7006$.

4.2 Recency Effect

Second, the influence of recency on adaptive video streaming is evaluated. Recency refers to the human brain's preference to attach higher importance to recent stimuli. However, the content duration has an influence on the occurrence of recency effects. However, the main intention of this study is to draw conclusions how to design proper adaptation algorithms. Therefore, we only consider the short video sequences without analyzing the the influence of content duration on the occurrence of recency effects.

In particular, it will be investigated if the video is perceived differently when it ends with high or low quality. Three switching patterns (N=1, 3, 5 switches) ending at high quality and their mirrored patterns ending at low quality were displayed to the test users. Figure 2 shows the comparison of the test videos. Again the MOS and 95% confidence intervals are plotted over the number of switches. Dark bars indicate a low quality ending, light bars a high quality ending. For three or five switches, the confidence intervals overlap and no impact of the video ending can be found. If the video contains only one switch, the high end condition is perceived slightly better than the low end condition. However, this might be due to the specific characteristic of this pattern which strictly separates high and low quality, containing 7.5s of low quality and afterwards 7.5s of high quality, or vice versa. Hence, it seems that the oscillation between high and low qualities removes the impact of the recency effect. In the following, we will focus on conditions with two or more switches that always end on the high quality level.

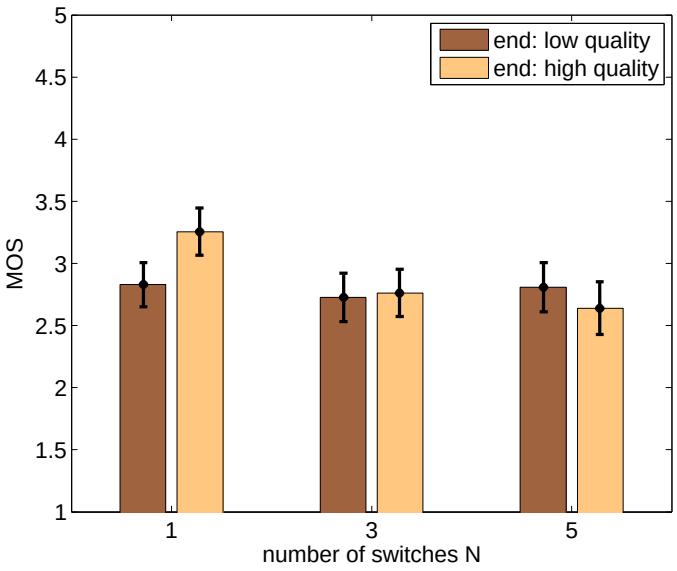


Figure 2: Recency Effect: Low vs High Quality Ending. Two-sample t-test does not reveal significant differences ($p = 0.2258$) between groups (ending low vs. high quality) with Cohen's $d = 0.1037$.

Another parameter of the recency effect is the recency time, i.e., the time how long high quality is played out after the last quality switch. Two switching patterns (N=2 and 4 switches) were chosen and shifted within the video, resulting in different recency times while preserving all other characteristics of the adaptation pattern. Figure 3 shows the MOS and 95% confidence intervals over different recency times in seconds. It can be seen that for both switching patterns (N=2 dark bars, N=4 light bars), the recency time does not influence the perceived quality. All MOS values range around 3 and the confidence intervals overlap. Thus, no significant effect of recency time could be observed.

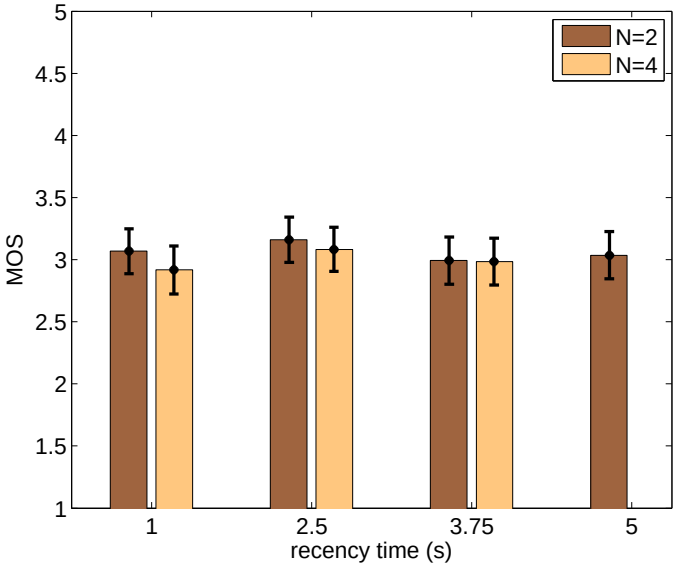


Figure 3: Recency Time Effect is not significant (see Table 2). Two-sample t-test does not reveal significant differences ($p = 0.3376$) between groups ($N = 2$ vs. $N = 4$) with Cohen's $d = 0.0693$.

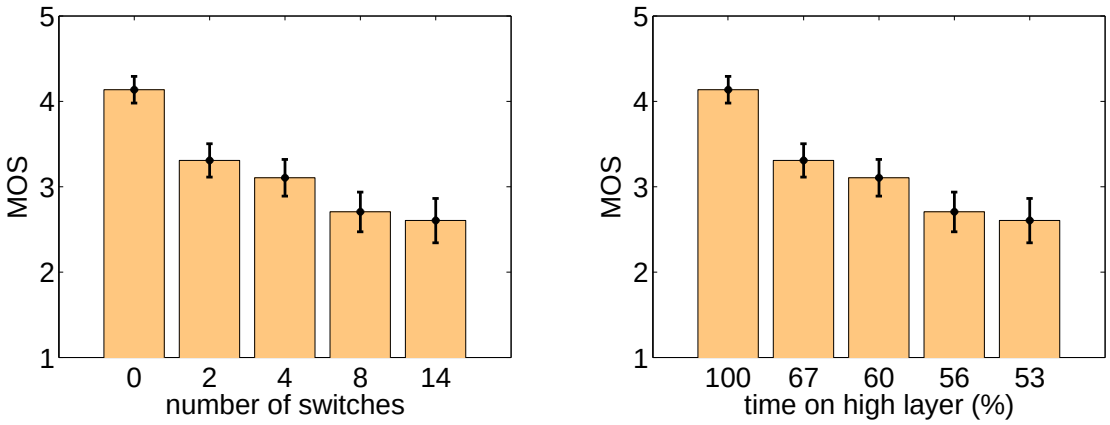
4.3 Frequency vs. Time on Highest Layer

Several works (e.g., [11, 18]) suggested an influence of the switching frequency, i.e., the number of switches, on the QoE of adaptive video streams. [19] showed that multiple quality switches are preferred over fewer switches, if the subject is able to view the high quality video for longer duration. However, in the results presented above (cf. Figures 1-3) the frequency effect seems negligible. Therefore, we revisited our experiments and found that not only the number of switches but also the corresponding time on the highest layer could be responsible for the observed effects.

In Figure 4a and 4b, the MOS and confidence intervals are plotted over the number of switches (a) and the corresponding time on the highest layer (b). Both parameters seem to have a significant effect on the perceived quality. Therefore, they have to be investigated separately in more detail towards a comprehensive QoE model.

For that, we designed two new experiments. First, the number of quality switches was varied whereas the time on the highest layer was kept constant. Figure 5 shows the MOS and 95% confidence intervals for different numbers of switches. The condition without switches reaches a MOS around 4. As soon as quality switches are present, it can be seen clearly that the number of switches has no significant impact on the QoE. For any number of switches the MOS ranges slightly below a value of 3 and the confidence intervals overlap.

In the second experiment the time on high layer was varied for two different numbers of switches (N=2 and 4). The results are plotted in Figure 6. The x-axis shows the time on the highest layer in percent and the y-axis shows the MOS and 95% confidence



(a) Number of switches **(b)** Time on highest layer

Figure 4: Frequency and time on highest layer changed simultaneously.

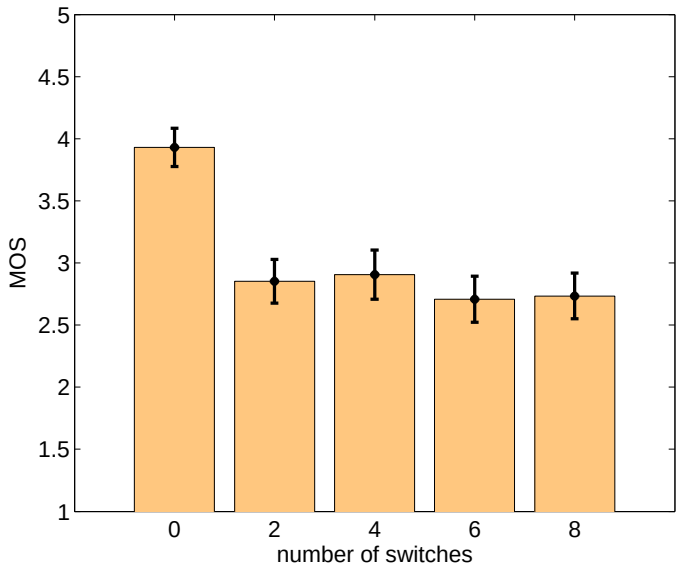


Figure 5: Impact of Switching Frequency. One-way ANOVA for this experiment returns the following p-values: a. $p = 0.0129$ for all conditions and b. $p = 0.7190$ for switches only, i.e. $N > 0$.

intervals. Again, the condition without switches (100% time on highest layer) reaches a MOS around 4. It can be observed for each pair of bars (N=2 dark, N=4 light) that the number of switches has no impact as the confidence intervals overlap. However, here an effect of the time on the highest layer is visible. The MOS decreases when the time on the highest layer decreases.

The results of related works might not be contradictory, as a quality switch typically implies a change of time on each layer. If quality switches are too frequent (i.e., user perceives only flickering), this might be worse than low video quality [11]. If quality

	r_s	F	p	η_p^2	f^2
Amplitude	-0.266	63.101	< 0.001	0.013	0.011
Last Quality Level	0.103	5.500	0.004	0.002	0.002
Recency Time	0.109	0.900	0.480	0.001	0.001
Switches	-0.221	1.736	0.139	0.001	0.001
Time on Highest Layer	0.295	17.742	< 0.001	0.043	0.037

(a) All configurations

	r_s	F	p	η_p^2	f^2
Amplitude	-0.302	66.350	< 0.001	0.017	0.015
Last Quality Level	0.018	2.508	0.082	0.001	0.001
Recency Time	0.100	0.635	0.638	0.001	0.001
Switches	-0.088	1.750	0.136	0.002	0.002
Time on Highest Layer	0.143	7.799	< 0.001	0.022	0.020

(b) Configurations with quality switches only

Table 2: Quantification of main effects based on one-way ANOVA.

switches are normally frequent, they will only point the user to a perceivable degradation/improvement. Once the user is aware of the degradation/improvement, again the duration matters. Thus, the impact of frequency is less than the impact of time on each layer. Similarily, [19] assumes that 'long low-quality video segments preceded by much higher quality segments evoke a strong negative response.'

4.4 Simplified QoE Model and its Consequences

In order to consistently revisit all investigated influence factors, statistical analyses of effects were conducted. Table 2 shows the effect sizes both for all configurations and for the configurations which contain quality switches only. The table shows in each row one of the investigated influence factors and its effect on the subjective quality ratings. The Spearman correlation coefficient r_s indicates how much the influence factor and the subjective quality ratings are associated. It can be seen that there is no high correlation for all factors. Second, a one-way analysis of variance (ANOVA) was conducted and the F-test statistic and the corresponding p-value are shown in the second and third column. The p-value indicates that both amplitude and time on highest layer are significant. Moreover, partial Eta-squared (η_p^2) and Cohen's f^2 were computed. The η_p^2 values indicate that there is a small effect for both amplitude and time on highest layer, which has also a small effect according to f^2. Having conducted these extensive statistical analyses, no effect can be observed for last quality level, recency time, and number of switches.

The results from the previous investigations yield implications for a new simplified

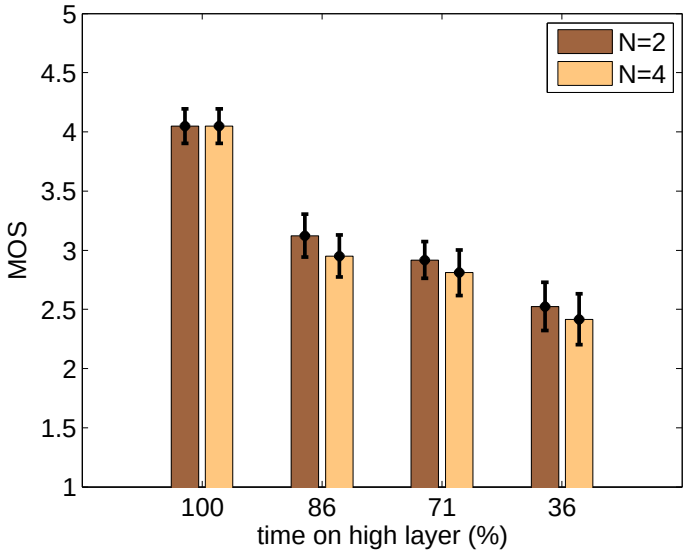


Figure 6: Time on Highest Layer is a main influence factor. One-way ANOVA for this experiment returns $p = 5.44e - 037$. Two-sample t-test does not reveal significant differences ($p = 0.1055$) between groups ($N = 2$ vs. $N = 4$) with Cohens' $d = 0.1398$.

QoE model. The proposed model is depicted in Figure 7 and takes into account the two main effects, amplitude and time on highest layer. The vertical coordinate is the MOS value prediction based on the time on highest layer, which is plotted on the x-axis. To account for the amplitude effect, the MOS value is bounded by quality $y_H = f(1)$ on highest layer and $y_L = f(0)$ on lowest layer, respectively. The gray bounds correspond to the mean value and 95% confidence intervals of the ratings of the video clip with constant high ([4.09;4.18]) or low ([2.37;2.66]) quality and were obtained in a separate experiment. The black data points correspond to the subjective ratings (mean and 95% confidence intervals) of the time on highest layer experiment (cf. Figure 6). Then, the relationship between MOS and time t on highest layer (%) follows the IQX hypothesis [20] $f(t) = \alpha e^{\beta t} + \gamma$. The data points fit the exponential model very well which is indicated by a high coefficient of determination $R^2 = 0.98$.

As the model only indicates the QoE for adaptation between two layer content, further subjective studies with more layers are needed in order to generalize the model. Following the above findings, these studies need to consider especially the amplitude and time on each individual layer, and have to investigate these influence factors in detail.

By having a QoE model which depends on amplitude and time on individual layer only, QoE monitoring becomes much easier. Both objective or subjective metrics can be used to assess the quality of each layer, and depending on the time on each layer, an exponential decay can be applied to obtain a QoE prediction. In a similar fashion, [21] reported a decent behavior of simple temporal pooling approaches, which showed a high correlation to subjective ratings of adaptive video streams. Here, a reduction of the monitoring complexity could be achieved by taking into account objective quality values

per layer (cf. amplitude factor) and aggregating them over the time on each specific layer (cf. time on individual layer factor).

Moreover, the knowledge of a simple QoE function can be used to model the optimal QoE. The approach presented in [22] can be adapted to incorporate a value function that takes amplitudes and times on individual layers into account. Thereby, optimal solutions for given network conditions can be obtained.

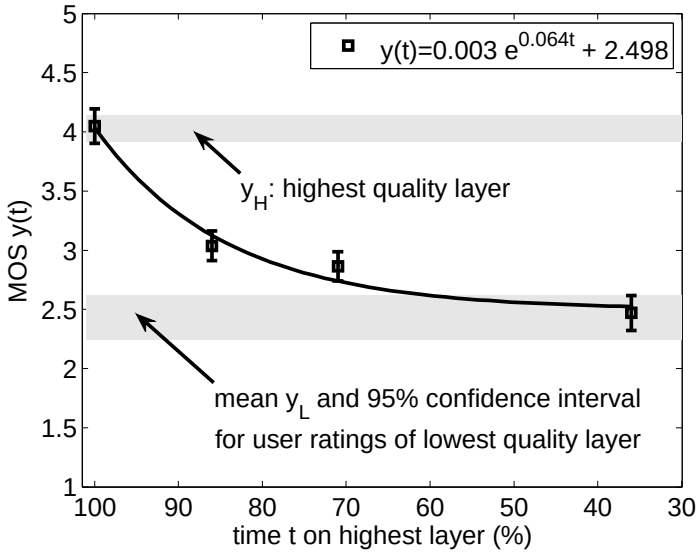


Figure 7: Simplified QoE model for two layers.

Finally, this lays the foundations for novel HAS algorithms which incorporate QoE management. The results of this work already provide the implication that the number of switches needs not be optimized in first place in future adaptation strategies. Instead, up-switching should be done as early as possible because it increases the time on highest layer. However, more research is needed in order to develop technical solutions which reach the optimal adaptation strategy.

5 Conclusions and Future Work

In this paper, the influence factors of the adaptation parameters for HTTP adaptive streaming were investigated and evaluated using subjective tests conducted in a crowdsourcing platform. As used test material we rely on a video clip of 14 seconds. The results clearly demonstrate the high impact of the switching amplitude between two played back representations, and that recency effects can be neglected if more than two switches occur. Further, it turned out, that the time on highest video quality layer has a significant impact on the QoE, and that the number of quality switches can be neglected. Based on these results, a simplified QoE model for adaptive streaming was derived. Further evaluations and surveys have to be conducted to fortify the findings and to allow the creation of a generally accepted QoE model for HTTP adaptive streaming.

6 Acknowledgements

This work was partly funded by Deutsche Forschungsgemeinschaft (DFG) under grants HO 4770/1-1 and TR257/31-1 and in the framework of the EU ICT Project SmartenIT (FP7-2012-ICT-317846). The authors alone are responsible for the content.

References

[1] Cisco Visual Networking Index, "Forecast and Methodology, 2012-2017 (29 May 2013)."

[2] C. Müller, S. Lederer, and C. Timmerer, "An evaluation of dynamic adaptive streaming over http in vehicular environments," in *4th Workshop on Mobile Video*, pp. 37–42, ACM, 2012.

[3] International Standards Organization/International Electrotechnical Commission (ISO/IEC), "23009-1:2012 Information Technology – Dynamic Adaptive Streaming over HTTP (DASH) – Part 1: Media Presentation Description and Segment Formats," 2012.

[4] International Telecommunication Union, "ITU-T Recommendation P.10/G.100, Amendment 2: Vocabulary and Effects of Transmission Parameters on Customer Opinion of Transmission Quality," 2006.

[5] R. Schatz, T. Hoßfeld, L. Janowski, and S. Egger, "From Packets to People: Quality of Experience as New Measurement Challenge," in *Data Traffic Monitoring and Analysis: From Measurement, Classification and Anomaly Detection to Quality of Experience* (M. M. Ernst Biersack, Christian Callegari, ed.), Springers Computer Communications and Networks series, Volume 7754, 2013.

[6] A. Finamore, M. Mellia, M. M. Munafò, R. Torres, and S. G. Rao, "YouTube Everywhere: Impact of Device and Infrastructure Synergies on User Experience," in *Internet Measurement Conference*, (Berlin, Germany), 2011.

[7] B. Gardlo, S. Egger, M. Seufert, and R. Schatz, "Crowdsourcing 2.0: Enhancing Execution Speed and Reliability of Web-based QoE Testing," in *IEEE International Conference on Communications (ICC 2014)*, (Sydney, Australia), 2014.

[8] C. Sieber, T. Hoßfeld, T. Zinner, P. Tran-Gia, and C. Timmerer, "Implementation and User-centric Comparison of a Novel Adaptation Logic for DASH with SVC," in *IFIP/IEEE International Workshop on Quality of Experience Centric Management (QCMan), Best Paper Award*, (Ghent, Belgium), May 2013.

[9] G. Ghinea and J. P. Thomas, "QoS Impact on User Perception and Understanding of Multimedia Video Clips," in *6th ACM International Conference on Multimedia*, (Bristol, UK), 1998.

[10] J.-S. Lee, F. De Simone, and T. Ebrahimi, "Subjective Quality Evaluation via Paired Comparison: Application to Scalable Video Coding," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 882–893, 2011.

[11] P. Ni, R. Eg, A. Eichhorn, C. Griwodz, and P. Halvorsen, "Flicker Effects in Adaptive Video Streaming to Handheld Devices," in *19th ACM International Conference on Multimedia (MM 2011)*, (Scottsdale, AZ, USA), 2011.

[12] K. D. Singh, Y. Hadjadj-Aoul, and G. Rubino, "Quality of Experience Estimation for Adaptive HTTP/TCP Video Streaming Using H. 264/AVC," in *IEEE Consumer Communications and Networking Conference (CCNC)*, (Las Vegas, NV, USA), 2012.

[13] M. Zink, J. Schmitt, and R. Steinmetz, "Layer-encoded Video in Scalable Adaptive Streaming," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 75–84, 2005.

[14] C. Keimel, J. Habigt, C. Horch, and K. Diepold, "QualityCrowd: A Framework for Crowd-based Quality Evaluation," in *Picture Coding Symposium (PCS), 2012*, pp. 245–248, IEEE, 2012.

[15] 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, Feb. 2014.

[16] C. Sieber, "Holistic Evaluation of Novel Adaptation Logics for DASH and SVC," master thesis, University of Würzburg, Germany, Aug. 2013. http://opus.bibliothek.uni-wuerzburg.de/.

[17] T. Hoßfeld and C. Keimel, "Crowdsourcing in QoE Evaluation," in *Quality of Experience: Advanced Concepts, Applications and Methods* (A. R. Sebastian Mller, ed.), Springer: T-Labs Series in Telecommunication Services, ISBN 978-3-319-02680-0,, Mar. 2014.

[18] B. Lewcio, B. Belmudez, A. Mehmood, M. Wältermann, and S. Möller, "Video Quality in Next Generation Mobile Networks – Perception of Time-varying Transmission," in *IEEE International Workshop Technical Committee on Communications Quality and Reliability (CQR 2011)*, (Naples, FL, USA), 2011.

[19] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. De Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 6, no. 6, pp. 652–671, 2012.

[20] M. Fiedler, T. Hossfeld, and P. Tran-Gia, "A Generic Quantitative Relationship between Quality of Experience and Quality of Service," *IEEE Network*, vol. 24, no. 2, pp. 36–41, 2010.

[21] M. Seufert, M. Slanina, S. Egger, and M. Kottkamp, "To Pool or Not To Pool: A Comparison of Temporal Pooling Methods for HTTP Adaptive Video Streaming," in *5th International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 52–57, IEEE, 2013.

[22] K. Miller, N. Corda, S. Argyropoulos, A. Raake, and A. Wolisz, "Optimal Adaptation Trajectories for Block-Request Adaptive Video Streaming," in *20th International Packet Video Workshop (PV)*, pp. 1–8, IEEE, 2013.

Appendix: Crowdsourcing Campaigns and Test Sequences

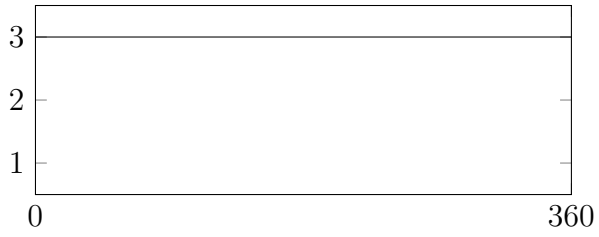
Table 3: Description of crowdsourcing campaigns reflecting individual user studies for testing several HAS influence factors as summarized in Table 1. The same identifiers for the campaigns C2–C5 are taken as in [16], while campaigns C6–C8 are subsequent experiments. The duration of a single test video sequence in C2–C5 is 15 s corresponding to 360 frames at a frame rate of 24 fps; the test sequence duration was set to 14 s in C6–C8 in order to investigate the influence of time on high layer and recency time. The column 'Begin-End' indicates whether the video playout begins/ends with high quality ('H') and low quality ('L'), respectively. The column '#Users' only considers reliable users passing the reliability checks. The column '#Seq.' shows the number of test video sequences which were shown to each user. The other columns give an overview on the quality switching pattern which are described in details in the corresponding Campaigns C2-C8.

Id	Amplitude	#Switches	Recency (s)	Begin-End	Time High (s)	#Users	#Seq.
C2	low	0,1,2,3,4,6,8	1.67 – 7.5	HH, HL	7.5 – 15	133	7
C3	high	0,1,2,3,4,8,14	1.00 – 7.5	HL,HH,LL	0 – 15	84	8
C4	high	0,1,2,3,4,5,7,8	1.67 – 7.5	HH, LH, LL	0 – 15	90	9
C5	high	0,1,2,3,4,5,6,8	1.50 – 7.5	HH, HL, LL	0, 7.5, 15	93	9
C6	high	0,2,4	1.00 - 5.0	HH	6.5, 14	112	8
C7	high	0,2,4	2.5, 3.75	HH	5, 7.5, 10, 14	140	7
C8	high	0,2,4	2.5, 3.75	HH	5, 10, 12, 14	90	7

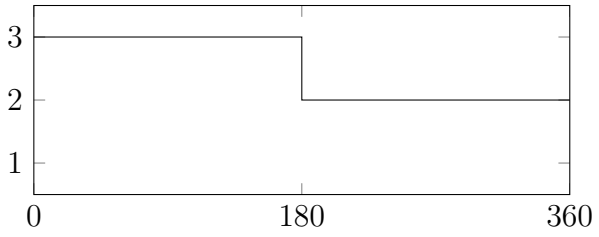
Table 4: Relation of crowdsourcing campaigns and the investigated effects by analyzing the corresponding test sequences over different campaigns. The individual quality switching patterns are depicted in the corresponding Campaigns C2-C8.

Id	Amplitude	Frequency	Recency Time	Begin-End	Time on Highest Layer
C2	b,c,d,e,g				
C3	b,c,d,e,f	a,c,e,g,h		b,d	a,c,e,f,g
C4				b,d,f	
C5		a,c,e,g,h		b,d,f	
C6			b,c,d,e,f,g,h		
C7		a,b,c,d,e,f,g			a,b,c,d,e,f,g
C8		a,b,c,d,e,f,g			a,b,c,d,e,f,g

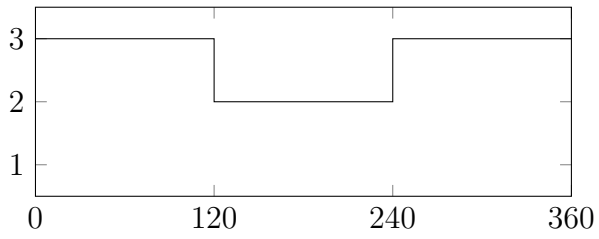
Campaign C2: Quality switching pattern with the amplitude level on the y-axis and the video time in terms of frames on the x-axis.



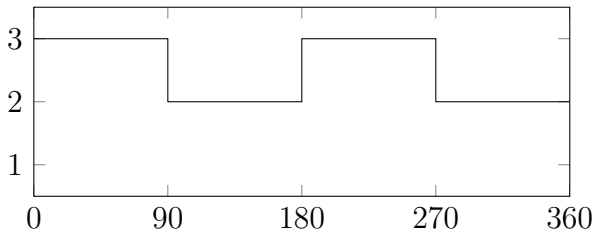
(C2.a)



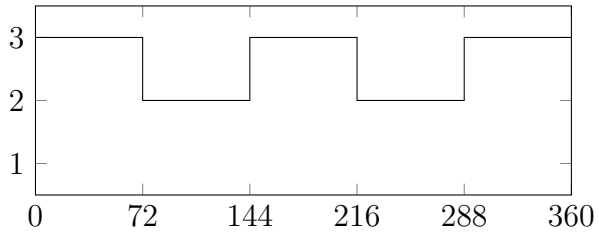
(C2.b)



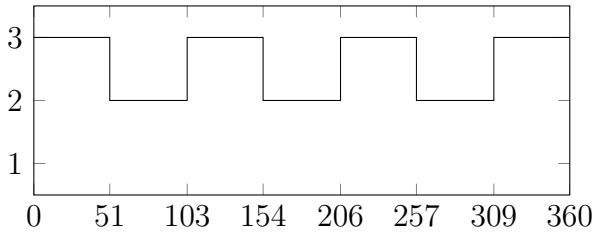
(C2.c)



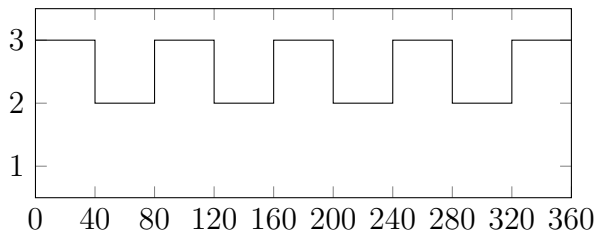
(C2.d)



(C2.e)



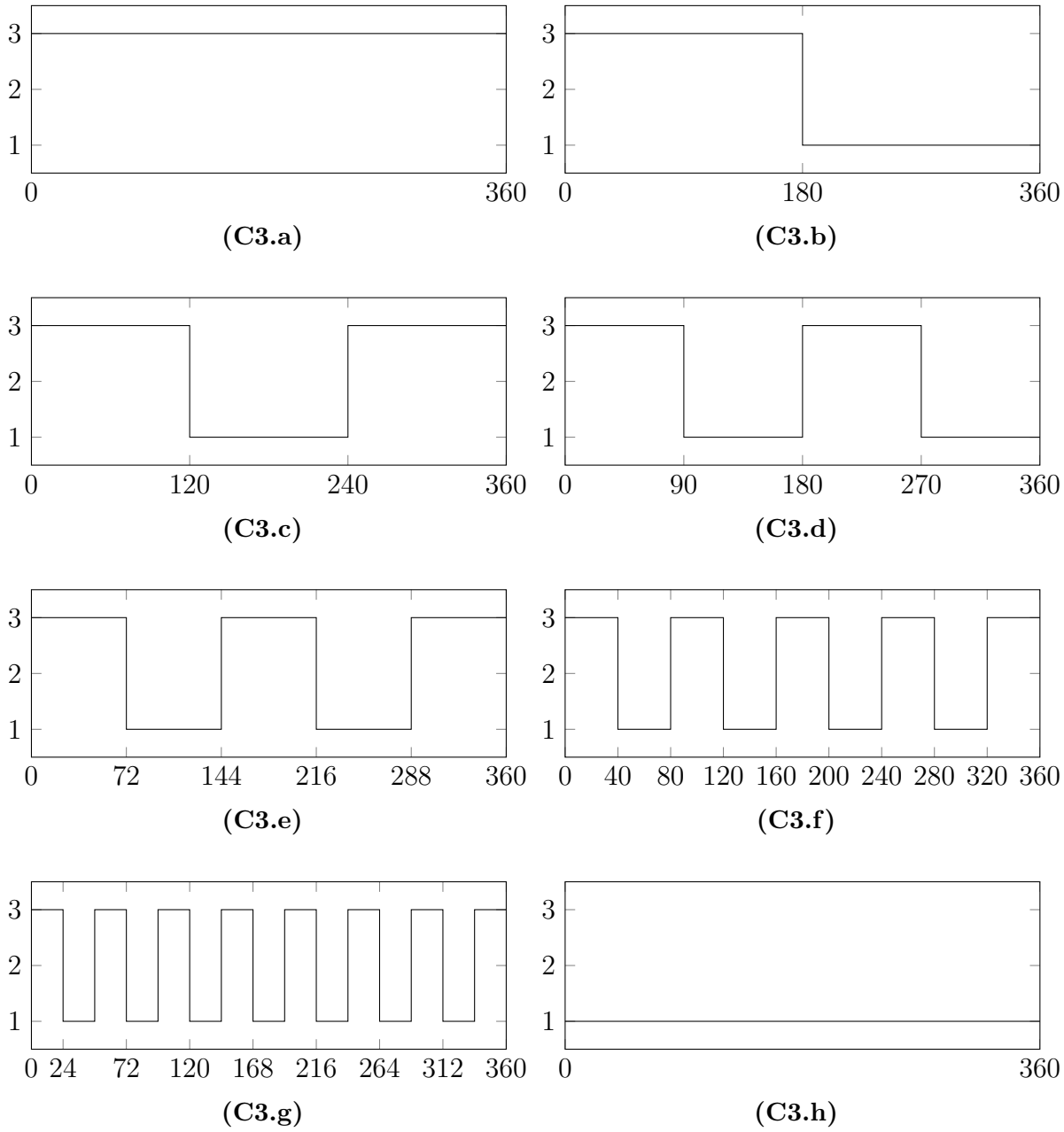
(C2.f)



(C2.g)

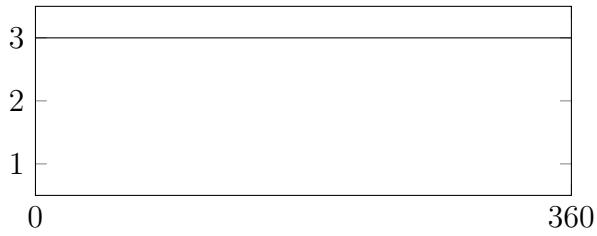
Id	#Switches	Time High	Recency
C2-a	0	15.00 s	15.00 s
C2-b	1	7.50 s	7.50 s
C2-c	2	10.00 s	5.00 s
C2-d	3	7.50 s	3.75 s
C2-e	4	9.00 s	3.00 s
C2-f	6	8.50 s	2.13 s
C2-g	8	8.33 s	1.67 s

Campaign C3: Quality switching pattern with the amplitude level on the y-axis and the video time in terms of frames on the x-axis.

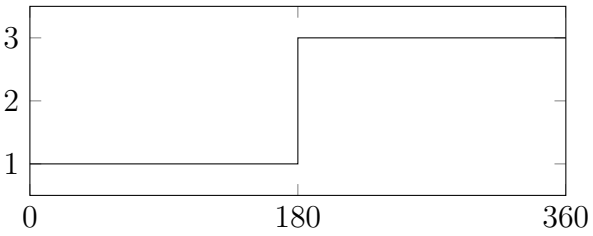


Id	#Switches	Time High	Recency
C3-a	0	15.00 s	15.00 s
C3-b	1	7.50 s	7.50 s
C3-c	2	10.00 s	5.00 s
C3-d	3	7.50 s	3.75 s
C3-e	4	9.00 s	3.00 s
C3-f	8	8.33 s	1.67 s
C3-g	14	8.00 s	1.00 s
C3-h	0	0.00 s	15.00 s

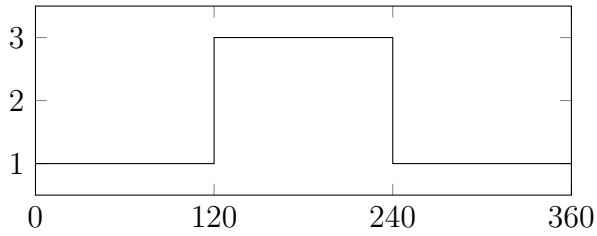
Campaign C4: Quality switching pattern with the amplitude level on the y-axis and the video time in terms of frames on the x-axis.



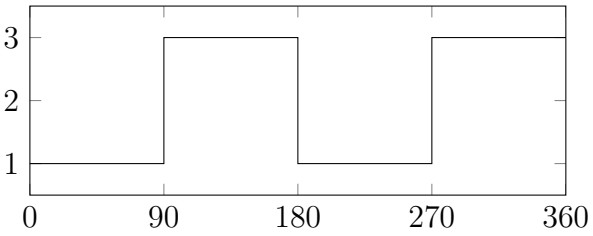
(C4.a)



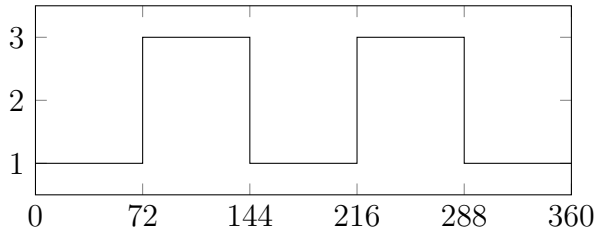
(C4.b)



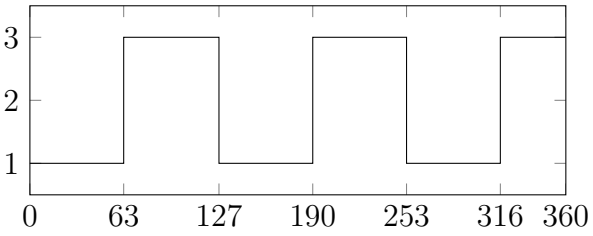
(C4.c)



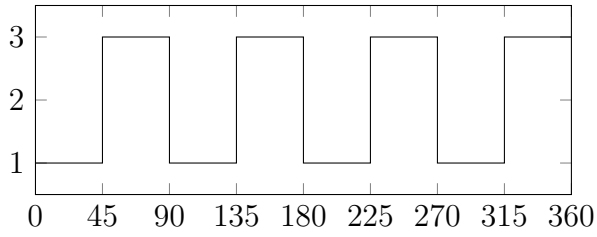
(C4.d)



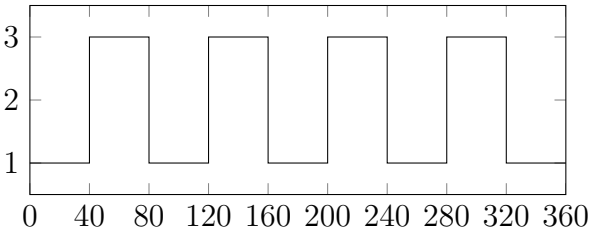
(C4.e)



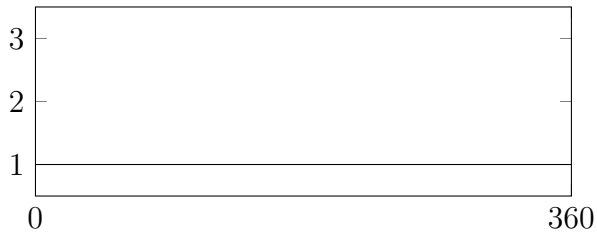
(C4.f)



(C4.g)



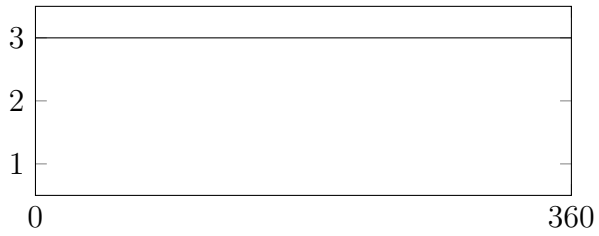
(C4.h)



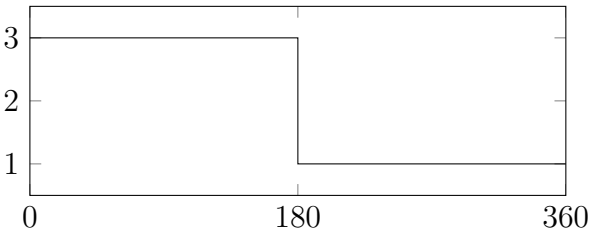
(C4.i)

Id	#Switches	Time High	Recency
C4-a	0	15.00 s	15.00 s
C4-b	1	7.50 s	7.50 s
C4-c	2	5.00 s	5.00 s
C4-d	3	7.50 s	3.75 s
C4-e	4	6.00 s	3.00 s
C4-f	5	7.13 s	1.83 s
C4-g	7	7.50 s	1.88 s
C4-h	8	6.67 s	1.67 s
C4-i	0	0.00 s	15.00 s

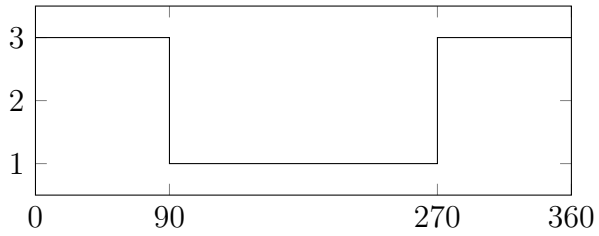
Campaign C5: Quality switching pattern with the amplitude level on the y-axis and the video time in terms of frames on the x-axis.



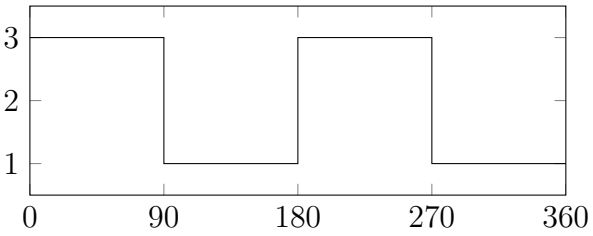
(C5.a)



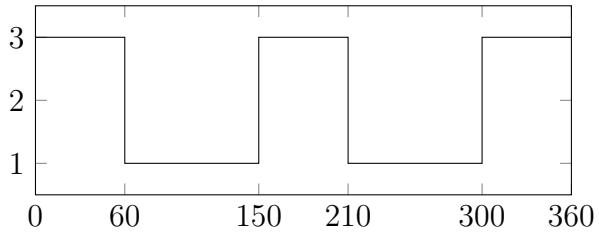
(C5.b)



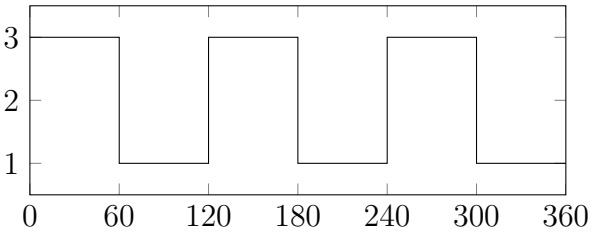
(C5.c)



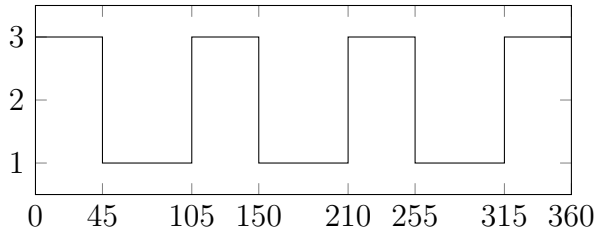
(C5.d)



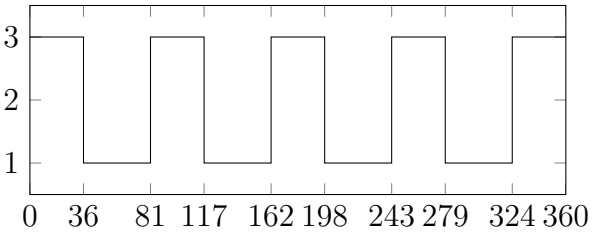
(C5.e)



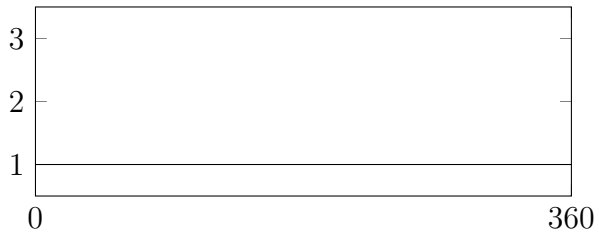
(C5.f)



(C5.g)



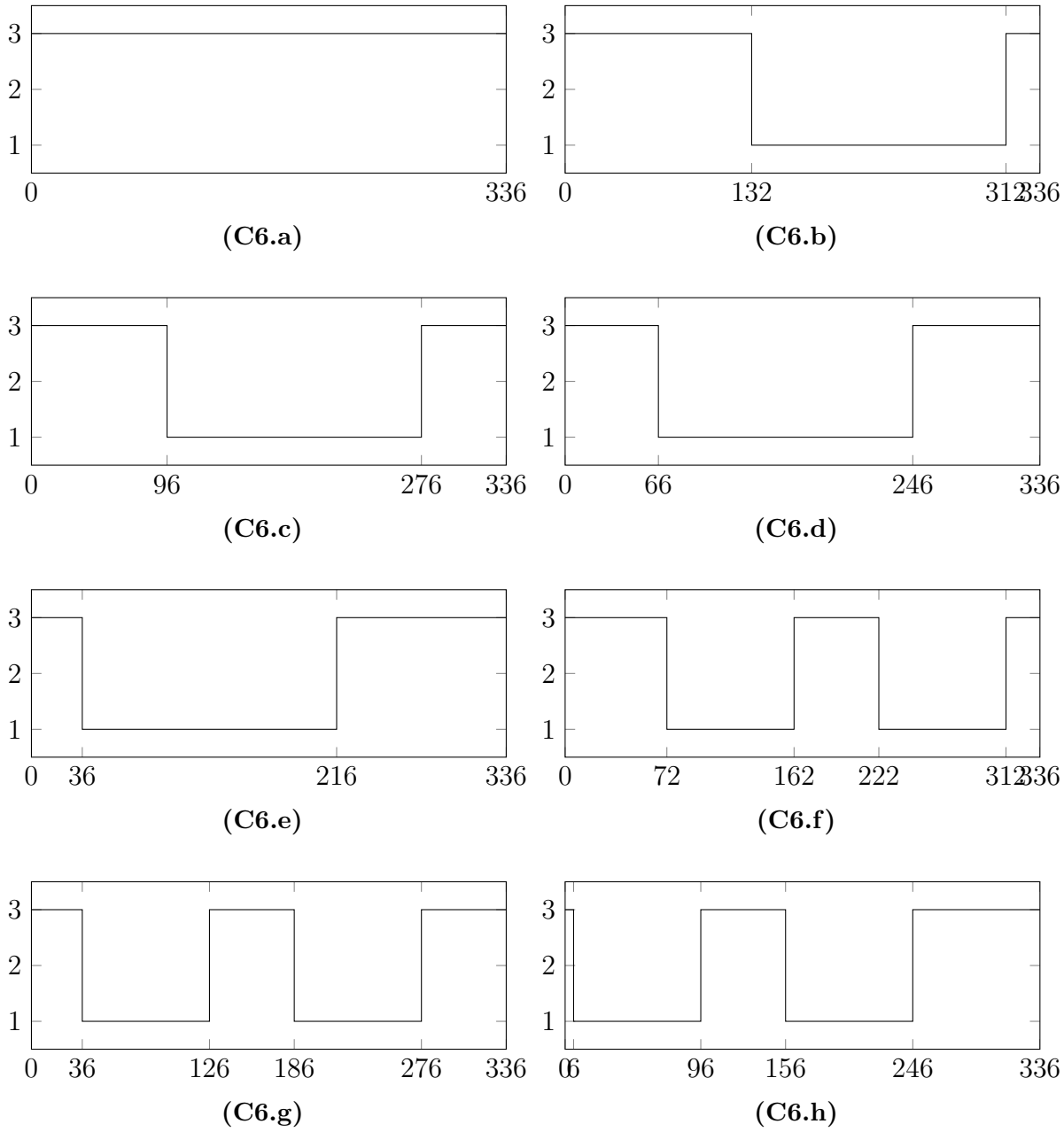
(C5.h)



(C5.i)

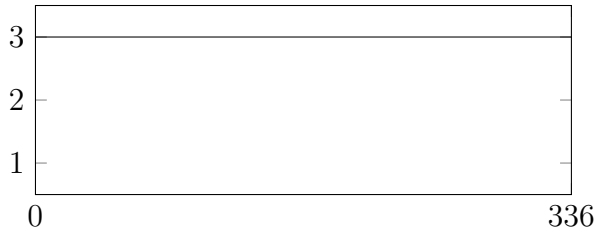
Id	#Switches	Time High	Recency
C5-a	0	15.00 s	15.00 s
C5-b	1	7.50 s	7.50 s
C5-c	2	7.50 s	3.75 s
C5-d	3	7.50 s	3.75 s
C5-e	4	7.50 s	2.50 s
C5-f	5	7.50 s	2.50 s
C5-g	6	7.50 s	1.88 s
C5-h	8	7.50 s	1.50 s
C5-i	0	0.00 s	15.00 s

Campaign C6: Quality switching pattern with the amplitude level on the y-axis and the video time in terms of frames on the x-axis.

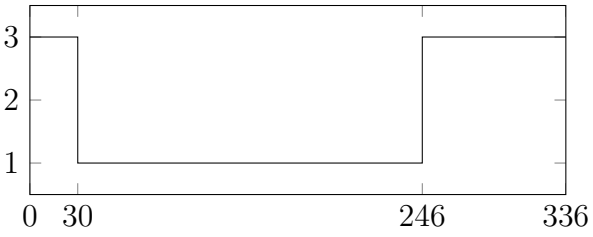


Id	#Switches	Time High	Recency
C6-a	0	14.00 s	14.00 s
C6-b	2	6.50 s	1.00 s
C6-c	2	6.50 s	2.50 s
C6-d	2	6.50 s	3.75 s
C6-e	2	6.50 s	5.00 s
C6-f	4	6.50 s	1.00 s
C6-g	4	6.50 s	2.50 s
C6-h	4	6.50 s	3.75 s

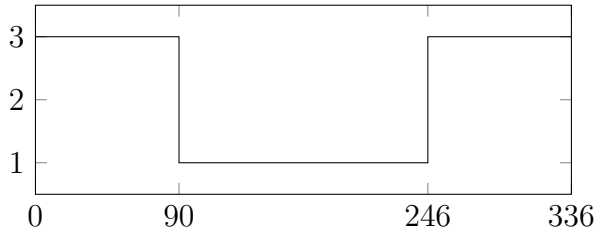
Campaign C7: Quality switching pattern with the amplitude level on the y-axis and the video time in terms of frames on the x-axis.



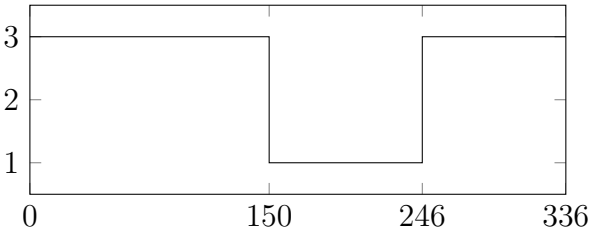
(C7.a)



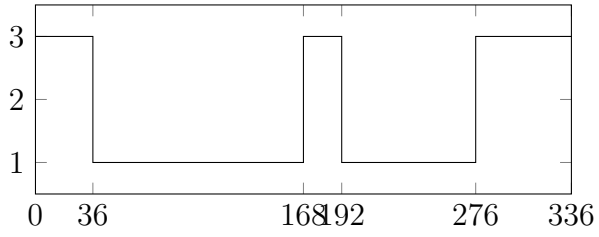
(C7.b)



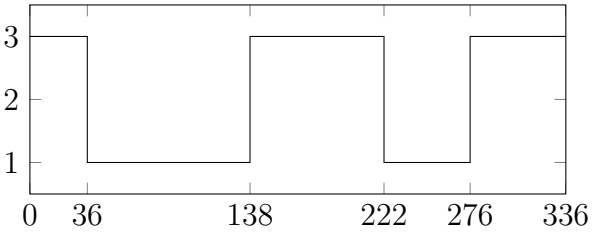
(C7.c)



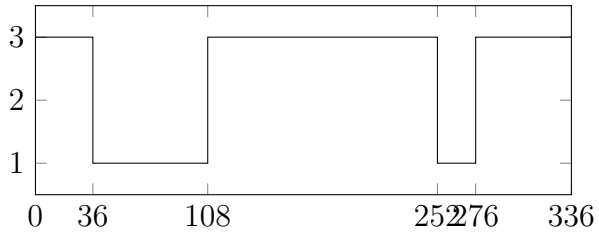
(C7.d)



(C7.e)



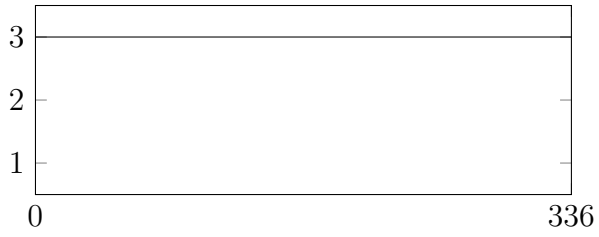
(C7.f)



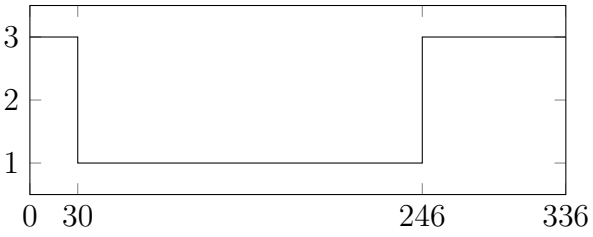
(C7.g)

Id	#Switches	Time High	Recency
C7-a	0	14.00 s	14.00 s
C7-b	2	5.00 s	3.75 s
C7-c	2	7.50 s	3.75 s
C7-d	2	10.00 s	3.75 s
C7-e	4	5.00 s	2.50 s
C7-f	4	7.50 s	2.50 s
C7-g	4	10.00 s	2.50 s

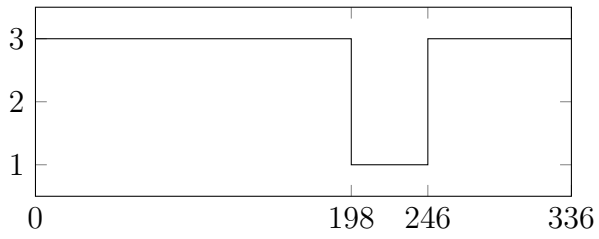
Campaign C8: Quality switching pattern with the amplitude level on the y-axis and the video time in terms of frames on the x-axis.



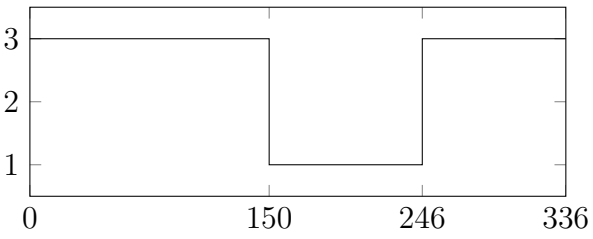
(C8.a)



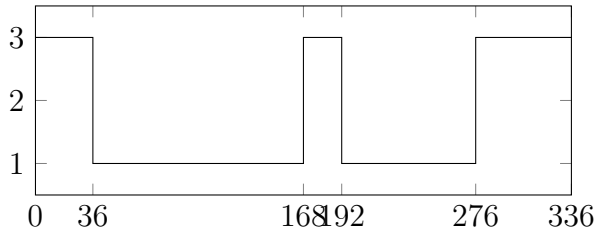
(C8.b)



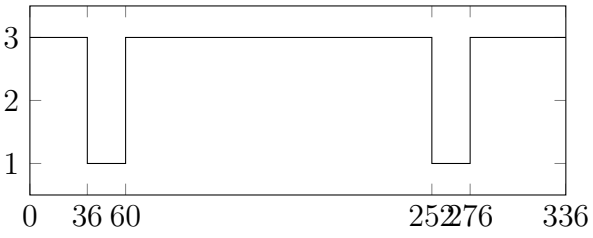
(C8.c)



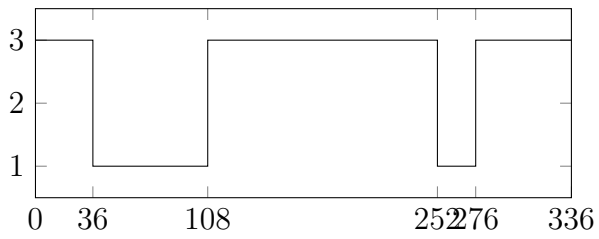
(C8.d)



(C8.e)



(C8.f)



(C8.g)

Id	#Switches	Time High	Recency
C8-a	0	14.00 s	14.00 s
C8-b	2	5.00 s	3.75 s
C8-c	2	12.00 s	3.75 s
C8-d	2	10.00 s	3.75 s
C8-e	4	5.00 s	2.50 s
C8-f	4	12.00 s	2.50 s
C8-g	4	10.00 s	2.50 s