Network Quality Differentiation: Regional Effects, Market Entrance, and Empirical Testability

Toni Mäki
VTT Technical Research Centre of Finland
Oulu, Finland.

Patrick Zwickl
University of Vienna
Faculty of Computer Science
Vienna, Austria

Martín Varela
VTT Technical Research Centre of Finland
Oulu, Finland.

***Abstract*—While for offline business models it is not seem necessary to reiterate the close relationship between quality and price, for Internet services the quality-based, i.e., Quality of Experience (QoE), and customer-centric pricing is non-trivial. As insufficient data exists today to successfully commercialise QoE, this paper collects the integral empirical Willingness-To-Pay (WTP) data for the case of online video services. This work reproduces and extends a previous study in two dedicated campaigns in Austria and Finland. The campaigns study QoE and WTP related to Dynamic Adaptive Streaming over HTTP (DASH). They also confirm or disprove previous studies, openly share the data, and provide empirical background information on the purchasing behavior of customers. Due to the testing at two locations, we can further first time study whether cultural or regional differences affect the purchasing behaviors of such services. Additionally this paper gives insights and updated methodological guidance on conducting future WTP studies.**

I. Introduction

The success of online services depends on several factors such as the value they provide (to match and satisfy customer demand), Quality of Experience (QoE), pricing strategies, but also on optimal use of resources (for cost efficiency). While the value proposition of such services may primarily be defined by the provided contents (e.g., video content to be streamed; quality of the videos), communication services can substantially affect the experience for customers. From the networking research and business point of view, the questions of quality, pricing, resource management and the interplay of these factors are, thus, the most interesting ones in understanding roles and co-operation of operators, ISPs, end customers and other stakeholders.

Service providers and operators have certain trade-offs to take into account when dimensioning for their service. They can try and minimize their costs, risking a lower-quality service, or they can try and offer the best possible quality to their users, with the risk of being inefficient in terms of cost (as achieving high quality levels in online services most often involves a significant investment in terms of resources). Between those extremes, there is of course a range of cost / quality ratios that can be planned for. QoE research gives good indications on managing such kinds of trade-offs.

However, optimisations are not only possible on the dimensioning side, but are also necessary for pricing and market strategy: While not very common today, service providers can make use of price discrimination based on quality, customer segment, regional factors. etc. In the context of QoE, the quality-based discrimination where operators offer pricing tiers with correspondingly different service quality levels are of outmost interest. Doing so in an optimal manner requires an understanding of how users perceive the value of the service and service quality (i.e., its *utility*), and how it translates to monetary means (i.e., revenues). This is a significantly different assessment to classical QoE testings as service and quality appeal may not equally translate to purchases or (high) WTP.

In contemporary markets the service offering and pricing can face highly dynamic competition as new challengers try to enter the field or existing companies try to increase their market share. This may also affect the users' opinions and their expectations on quality and pricing. Therefore it is important to understand and be able to estimate what kind of effects (sometimes necessary) tariff changes may incur. Additionally, for a new service or company it is important to plan the market entry properly. Market entrance pricing is a key element in this planning. While low entrance pricing may attract users, the later increases may prove to be difficult to implement, renderering the business unsustainable.

In this paper we propose to address the question of how users perceive the value of better quality in an online video service (*à la* Netflix), by means of an experiment on their Willingness-To-Pay (WTP) for different quality levels, as a close metric to utilities for ISPs. We further investigate how the relation of QoE and WTP is affected by different tariff changes and cultural or regional effects. The work presented herein replicates and expands upon a previous work in [1], where the problem was systematically studied. The present work differs from the previous approach by using an entirely paperless test laboratory, but also recent codec, i.e., H.265 / HEVC, and video adaptation advancements, DASH. The present work can also be considered to be a retesting of the results in [1], with target of the trial data to be openly accessible for the research community, which is not the case with the previous results. The experiments were carried out with almost identical setups in two labs, at the University of Vienna, in Austria, and at VTT, in Oulu, Finland. This allows the unique comparison of regional effects that may affect the utility and, thus, WTP for network video quality services.

From the results of the earlier work [1] we can isolate the following null hypotheses that were studied in this work:

Hypothesis 1 *WTP for network video quality upgrades does not exist.*

Hypothesis 2 *Historic pricing does not affect the market entrance of quality enhanced network video services.*

Hypothesis 3 *Different consumer segments do not make different quality - price decisions.*

The execution of the similar campaigns in two countries allows for testing possible variation in WTP between the two cultures. Additionally, it is known that the consumer prices (in relation to purchasing power) in Finland are higher than those in Austria (Comparative price levels 123.2 and 106.8, respectively[1]). Therefore we postulate the following null hypothesis:

Hypothesis 4 *WTP for quality-differentiated network video services is not affected by regional or cultural factors.*

The remainder of this work is structured as follows: in Section II we cover the relevant related work. The experiment environment, design and both setups are described in Section III before presenting the results in Section IV and analysis of key findings in Section V. The paper is finished with conclusions in Section VI.

II. Related Work

QoE has been a vital research topic in telecommunications for years. Especially the empirical perspective, both laboratory and field, to map technical the QoS to a subjective representation of QoE has received substantial attention [2], [3]. Numerous standards and recommendations, e.g., [4] and [5], have improved the test practices in order to obtain the reproducible, consistent results. The transfer of empirical or estimated QoE data to the provisioning of network resources has been, for example, discussed in [6] and more access-oriented in [7]. Despite the usefulness of such data and practices, the economic utilization has been hampered by several knowledge gaps:

1) The mapping of QoE to purchasing or spending behaviours
2) Communication problems [8] due to the experience product nature [9] of network quality
3) Difficult generalisation of data across individual measurements [10]

The most pressing issue is the first one listed above, as QoE information needs to be transferred to perspective of business models: utilities, product demand, etc. While an early work [11] has targeted the assessment of WTP for network quality the community has been silent for years afterwards. In the last few years this problem was then finally targeted from several perspectives:

1) The fixed-point model of QoE [12] which formalises the interaction between price and subjective quality experience
2) Empirical confirmation of early WTP results in [13] and [1], as well as the exploration of QoE spending phenomena, e.g., related to cognitive dissonance [14]
3) Approximation of WTP from QoE and other results in [10]

The recent empirical efforts for understanding WTP have focused on careful laboratory setups by learning from the experiences in QoE testing. Contrary to the approach in [11], [13] and especially [1] have strictly moderated the information that is provided to the user. In other words, these studies have reduced the usage complexity and eliminated several biases, such as an inherent convergence to the mean (of the quality range) effect. While [13] has tested UDP video transmissions under packet loss, [1] has used more modern adaptive streaming technologies based on TCP. Both studies were able to illustrate a reasonable WTP for enhanced network services, a clear trade-off management of subjects between quality and price concerns, and effects induced by historic pricing (i.e., "market entrance pricing" recommendations). Despite the promising results, the results of these studies are not openly available and due to the low sample sizes a confirmation of the effects is advisable. This work will, hence, bring the test design used in [1] to 2015 by conducting a new campaign using up-to-date codecs and video adaptation techniques.

III. Experiments Description

A. Overview

The tested scenario was about watching typical video streaming content in a living-room like environment and making video quality purchasing decisions. In addition to the hypotheses to be verified or disproved by the empirical laboratory-based studies, the work had also some technical and generic goals. Recent developments in multimedia technologies called for considering them also in WTP studies (in addition to numerous QoE studies covering them). The technological advancements compared to previous studies are summarized in Section III-B.

B. Technological Advancements & Changes

While the technical setup followed the initial testing in [1], a series of changes and advancements were necessary in order to meet the state-of-the-art of technologies and to respond to insights from the earlier tests.

In a way analogous to the initial testing, but contrary to the older studies such as [11] and [13], our study used adaptive streaming over TCP to allow dynamically applying quality changes and also to match the contemporary typical video usage. The standard DASH [15] was used as video streaming technology, instead of Apple's HTTP Live Streaming (used in [1]). The DASH content was played out for viewing with

[1]Eurostat, Purchasing Power Parities: http://ec.europa.eu/eurostat/web/purchasing-power-parities/ last accessed: December 11, 2015

the GPAC client[2] for Linux. GPAC player was manipulated in order to reduce the buffering and associated quality switching times, which was important for the experimental setup.

In earlier tests, the H.264 [16] video coding format was used. In the described campaigns, the substantially improved H.265 [17] (also referred to as HEVC [18], [19]) encoding was used instead. In pilot tests (executed prior to actual user campaigns), a reduced bandwidth demand of approximately 30% was witnessed in order to obtain comparable QoE values.

Contrary to a separate monitor in the initial testing, an iPad tablet computer was used to display both the available content (video library) and the price of the current selection (plus the information about the remaining reward user has).

Finally, while in the initial testing 3 test groups were used, the test group design was simplified in our approach, as sketched below.

C. Experiment Environment and Contents

1) Testing Environment: The subjective tests were executed in laboratory environment adhering the ITU-T P.910[4] (e.g. sample size and lighting conditions) and ITU-R BT.710 [5] (e.g. viewing distance) as closely as possible. The experiment set-up and how test subjects viewed and controlled the testing application is illustrated in Fig. 1. The main components of the test environment are described in Table I.

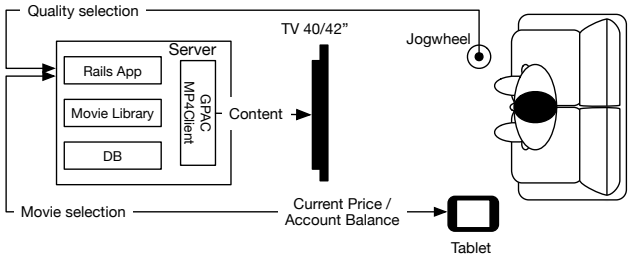

Fig. 1. Experiment set-up

2) Tested Contents: The contents were prepared by extracting the content from the purchased Blueray discs into full quality versions (in M4V and Matroska containers). Then the 20 minute clips of each movie were carefully selected and edited in full quality. Finally, the full quality clips were transcoded to different degraded qualities with help of *x265*[3] and *FFmpeg*[4] tools. The actual qualities (defined in terms of bitrate) are specified in upcoming sections. The original contents available in the Movie Library of the campaigns are listed in Table II. Some videos were offered in English and German in the Vienna trial.

[2]GPAC Multimedia Open Source Project: http://gpac.wp.mines-telecom.fr/, last accessed: December 11, 2015

[3]http://x265.org/

[4]https://www.ffmpeg.org/

TABLE I
TEST ENVIRONMENT: MAIN COMPONENTS

Component	Description	Function
Server	Ubuntu 14.04 LTS PC	Hosts the software application components used in the trial.
Rails App	Rails application	Implements the control UI used via tablet and also the control logic of the test.
Movie Library	File system	Selection of movies and descriptions accessible for Rails App and MP4Client
DB	SQLite3 Database	Holds the metadata and the results of the tests.
GPAC MP4Client	Media player application	Presents the DASH content to the subjects.
TV 40/42"	TV set	The screen used by subjects to view the content.
Jogwheel	Jogwheel device	Remote control device used to select the video quality.
Tablet	Tablet computer	iPad2 device used to read instructions, answer queries, select the movies and to view the currently selected quality's price and the current deposit balance.

TABLE II
THE CONTENT AVAILABLE IN THE MOVIE LIBRARY

Content	Description	Campaign
Grand Budapest Hotel	Comedy, movie, 2014.	Vienna (due to outliers removed)
Breaking Bad	Crime drama TV series, 2012.	Both
The Dark Knight Rises	Superhero movie, 2012.	Both
Edge of Tomorrow	Science fiction, 2014.	Both
Guardians of the Galaxy	Superhero movie, 2014.	Both
Harry Potter and the Order of the Phoenix	Fantasy movie, 2007.	Both
Inception	Science fiction movie, 2010.	Both
Interstellar	Science fiction movie, 2014.	Both
Oblivion	Science fiction movie, 2013.	Both
Oblivion	Science fiction movie, 2013.	Both
Orphan Black	Science fiction TV series, 2013	Both
The Hobbit: An Unexpected Journey	Fantasy movie, 2012	Both
Transcendence	Science fiction movie, 2014	Both (different edit)
Toy Story	3D animation movie, 2010	Oulu

D. Test Design

During the tests the users could choose from eight quality classes, where the quality of each class was controlled in terms of bitrate. The different quality levels are named $Q_0 \dots Q_7$, where Q_0 denotes the class with the lowest and Q_7 the class with the best quality.

The test design used three tariffs A, B and C with linear price curves from 0 to the resp. maximum prices p_{max} of €2, €3 and €4:

$$A := \{p_0 = 0, p_1 = 0.286, \dots, p_6 = 1.714, p_7 = 2\} \quad ,$$
$$B := \{p_0 = 0, p_1 = 0.429, \dots, p_6 = 2.571, p_7 = 3\} \quad ,$$
$$C := \{p_0 = 0, p_1 = 0.571, \dots, p_6 = 3.429, p_7 = 4\} \quad .$$

The experiment process is illustrated in Fig. 2. First, each subject received €10 at the beginning of the trial – the money was shown as a deposit on the screen and was initially provided symbolically in cash. The subjects could (but needed not) use this money to finance quality upgrades during the trial (€10 were sufficient to constantly watch the best quality).

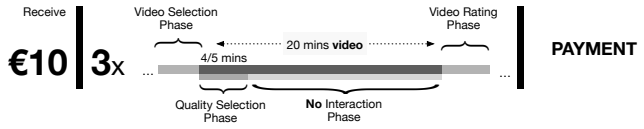

Fig. 2. Experiment sequence

After the trial, the remaining money on the deposit was to be paid out in cash, i.e., up to €10 could be paid out to the users in cash.

The actual experiment consisted of three measurements t_1, t_2, and t_3, each consisting of a 20 minutes video of the subject's choice and some ratings. Each user was assigned into *Group 1* or into *Control* group. The difference between the groups was in the tariffs (A, B, and C) that user was exposed to:

- *Group 1*: $t_1 : A \rightarrow t_2 : B \rightarrow t_3 : C$
- *Control*: $t_1 : B \rightarrow t_2 : B \rightarrow t_3 : C/A$

In other words, *Group 1* tested the *increasing* prices and the *Control* group had the stable pricing in t_1 and t_2 for comparison reasons. The *Control* group was further divided into two subgroups regarding the tariff of t_3 for a broader tariff comparison. Contrary to the initial trial, the decreasing prices were not tested, due to sample size reasons (in both the initial and retested trial) and the higher effects that have been witnessed for price increases in [1]. This test design allows within-subject comparisons, which require lower sample sizes for providing expressive results.

Analogously to the notions used in [13], each measurement t consisted of four phases (illustrated in Fig. 2):

1) **Video Selection Phase (VSP)**: The subject browsed our extensive library of modern video material and selected the content of her liking. The next phase was triggered upon the selection of the video.
2) **Quality Selection Phase (QSP)**: During the first 4 minutes (5 in Oulu trial) of the video watching the subject could freely test any quality level and evaluate the different quality-price tradeoffs (price was shown; quality was only perceived). When the QSP closed the quality level was fixed (to current one) and the price was finally deducted from the subjects balance.
3) **No Interaction Phase (NIP)**: The rest of the video was shown using the quality class selected by the user in the QSP. No further quality selection interaction was possible.
4) **Video Rating Phase (VRP)**: After the video had finished, the subject rated the QoE on the ACR-5 scale (*Bad*, *Poor*, *Fair*, *Good*, *Excellent*) and answered a binary acceptance question.

At the beginning, there was a pre-session questionnaire; the user was asked to specify her/his gender, age, education, Internet usage, Internet video purchasing habits and whether the user subscribes to some video services.

In a post-session questionnaire subject anwered the question "Did it feel like spending your own money?" in order to understand the validity of the test methodology. The users were also asked if they liked the available content.

E. Vienna Campaign

Using members of our faculty, several full-length pilot tests were conducted. Such tests served the elimination of test biases (such as unclear user interfaces), and assured the technical functioning of the trial and the meaningful parameterisation of the trial. Our expert users made the following noteworthy observations:

- The system is easy to use, the video content is interesting, and the scenario is realistic.
- Relative to H.264, H.265 performs surprisingly well with moderate bitrates
- Some experts did not recognise any quality gains above Q_4
- As the sound quality was rated to be insufficient, a new surround sound system was installed after the pilot test.

The actual campaign was conducted in our living laboratory in Vienna in July 2015. Twenty-two (22) test subjects completed all stages of the experiment with an average duration of $\approx$ 1.5 hours. Due to the three measurements, within-subject comparisons are enabled and 66 data points (purchases and associated QoE ratings) are available. Nine (9) subjects ($\approx 41\%$) were female and 19 had graduated from a university (typically with a master's degree or equivalent). The subjects belonged to the following age groups: two (2) subjects were between 10 and 19 years old, 11 were between 20 and 29, 6 were between 30 and 39 and 3 subjects were older. Their experiences with VoD services were limited, i.e., 11 subjects seldom purchased contents, one (1) did so weekly, and 7 had one or more video service subscriptions.

F. Oulu Campaign: Differences to Vienna Campaign

Some technical improvements were implemented by the recommendations derived from Vienna trial:

1) One of the observations was that HEVC provides good quality already on rather low bitrates. While generally a positive development, the high efficiency of the codec decreased the width of active decision making area in the trial as most users reached acceptable quality levels already on Q_3 or Q_4. To this end, the set of available quality levels were changed by adding more lower bit rate alternatives (See Table VI). The bit rates were selected so that their Peak Signal to Noise Ratio (PSNR) increases linearly. Also one of the videos was changed to start from beginning (Transcending) and one video was added to the library (Toy Story 3).
2) The price was made more prominent in control device by showing it in red colour.
3) The segment length of DASH content was set to 1 second (using also GOP size of 1 s). In the Vienna trial it was 2 seconds.

4) A 5 minute Video Selection Phase was used (as opposed to 4 minutes in the Vienna trials).

The test campaign was executed in November - December 2015 in the (living room-like) QoE laboratory at VTT premises in Oulu. Nineteen (19) subjects completed the experiment with average duration of $\approx$ 1 hour and 15 minutes. In this trial the three measurements led to 57 data points (purchases and QoE ratings). Most of the participants were VTT employees, and four (4) of them were female ($\approx$ 21%) and seventeen (17) held a university degree. The subjects belong to the following age groups: two (2) subjects were between 10 and 29 years old, 10 were between 30 and 39, five (5) between 40 and 49 and two were older. They used Internet rather much as eight (8) persons reported over five (5) hours daily usage and 10 the usage of 1 to 4 hours per day. Six (6) participants reported never buying video content from Internet, while the remaining participants used to purchase content seldomly. Five (5) participants reported buying their video content in HD/4K quality. Eleven (11) participants subscribed to some video services, but only two (2) persons subscribed to more than one service.

Due to the taxation laws of Finland, the reward could not be given out in cash (as intented). Instead the participants were rewarded with a movie ticket and variable amount of candy (depending on their deposit balance) they could select themselves. Nevertheless, the participants were led to believe in the beginning that they would receive the €10 as in the original design.

IV. Results

A. Vienna Campaign Results

TABLE III
2015 trial (Vienna): Quality levels Q_0 to Q_{17} in kBit/s.

Q_0	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7
128	256	512	1024	2048	4096	8192	16384

In the Vienna trial, the video qualities shown in Table IV were tested for quality levels Q_0 (poorest) to Q_7 (best). For these values the WTP shown in Table IV was observed.

TABLE IV
Spending per Tariff (*2015 trial, Vienna*).

	Overall	**Tariff A**	**Tariff B**	**Tariff C**
Median	€1.29	€0.86	€1.71	€1.71
(% of maximum)	(43%)	(43%)	(57%)	(43%)
Std. deviation	€0.87	€0.43	€0.80	€1.16
(% of maximum)	(26%)	(22%)	(27%)	(29%)

In other words, a substantial WTP for enhanced network video services was observed (median: €1.29), comparable to the experiences made in [1]. The majority of the subjects selected intermediary quality levels (see Fig. 5). Contrary to

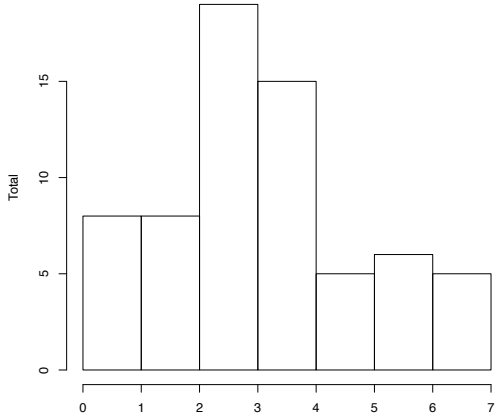

Fig. 3. Selected quality class Q_x (x-axis) in Vienna trial

the trial reported in [1], no peaks towards the range extrema were observed. This may be explained by the chosen codec: using the modern H.265 codec, the QoE saturates quickly with the chosen bitrates—H.265 provides better than expected QoE improvement for low bitrates. The codec starts to perform very well already at moderate bitrates, which yields surprisingly high QoE ratings, as Mean Opinion Score (MOS) on ACR-5 scale — see Fig. 4. Hence, subjects may not have perceived any quality difference for qualities better than Q_4, which distributed the premium segment between Q_5 and Q_7. In a restesting, we recommended a finer-grained quality offering in the lower to medium QoS range.

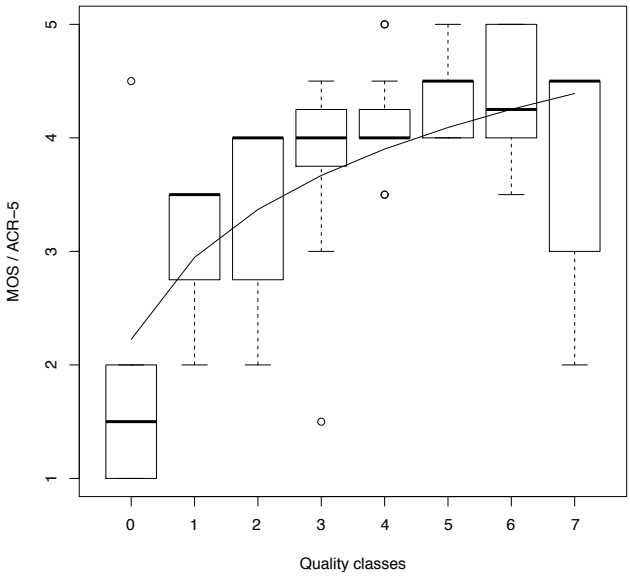

Fig. 4. Box plot of MOS ACR-5 ratings across all tariffs with logarithmic fit (*Vienna*).

The obtained WTP data was further relatively noisy —

high variation in t_1, low correlation between t_1 and t_2 in the control group. This may indicate that the video files offered in the marketplace were too heterogenous to allow a direct comparison. Thus, we recommend a further improvement of the carefully selected video library to assure even higher consistency. As especially the first measurement t_1 was very noisy, we suggest a longer QSP duration – in the Vienna trial only 3 minutes were used in pre-trial testing, which was later on extended to 4 minutes for the actual trial.

Due to the noisy data, the detailed analysis of the used groups has not been conclusive. The high noise in t_1 affects the comparison of t_2 across groups (equal tariffing in t_2, but unequal historic tariffs). However, when focusing on t_2 and t_3 of the control group, the subgroups with tariff sequences $B \rightarrow A$ and $B \rightarrow C$ can be compared.

TABLE V
SPENDING AS PERCENT OF p_{max} IN t_2 AND t_3 PER CONTROL SUBGROUP IN VIENNA TRIAL.

	t_2	t_3
	B: $p_{max} = 3$	**A:** $p_{max} = 2$
Mean	43%	45%
Median	43%	43%
	B: $p_{max} = 3$	**C:** $p_{max} = 4$
Mean	51%	29%
Median	57%	43%

As shown in Tab. V, the normalised expenditure is substantially affected by price increases, while price drops are hardly felt. When applying an ANOVA RM ($\alpha = 0.05$) to both the absolute and normalized spending, no significant time, group and group-time effects can be observed, however. This is caused by the test design that focused on the comparison of results in t_2 rather than t_3 and across test groups rather than looking at subgroups of the control group. Due to the high noise, not explained by rationales of subjects or provided feedback, especially in t_1, the analysis of historic pricing effects from [1] cannot be repeated for $t_1 \rightarrow t_2$.

An improved retesting shall target the working out of such effects in t_1 and t_2 or in a redesigned later test phase. The latter results highly correlate to the observations in [1]. Subjects seem to avoid a redecision in the case of price decreases, as not absolutely necessary, but immediately respond to price increases — see relationship to cognitive dissonance in [14].

Almost all subjects liked the provided contents (91%) and rated the "own money" feeling with 2.9 (ACR-5) on average (median: 3.0; "Fairly"). These results support the general functioning of the campaign design.

B. Oulu Campaign Results

The video qualities shown in Table VI were tested in Oulu trial. The overall and per tariff WTP are shown in Table VII. Interestingly, the WTP is higher than in the Vienna trial (€1.71). The results of both trials disprove Hypothesis 1.

TABLE VI
2015 TRIAL (OULU): QUALITY LEVELS Q_0 TO Q_{17} IN KBIT/S.

Q_0	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7
128	180	280	440	800	2548	8000	15000

As illustrated in Fig. 5 most of the users again selected intermediary quality levels. The peaks at the range extrema observed in the trial reported in [1] are present (esp. on the high-quality end), but they are not very pronounced. Still, the user groups present in [13] – (*price focused users*, *average users*, and *quality focused users* – can also be spotted in Fig. 5, which disproves the Hypothesis 3.

Fig. 6 illustrates the MOS of each quality level and it demonstrates the logarithmic nature of QoE. The implemented changes of the Oulu over the Vienna trial have caused more quality differentiation (by users) in medium and high bit rates, which is likely to have affected also the selected qualities of Fig. 5.

TABLE VII
SPENDING PER TARIFF (*2015 trial, Oulu*).

	Overall	**Tariff A**	**Tariff B**	**Tariff C**
Median	€1.71	€1.43	€1.71	€2.29
(% of maximum)	(71%)	(71%)	(57%)	(57%)
Std. deviation	€0.87	€0.43	€0.80	€1.16
(% of maximum)	(25%)	(18%)	(25%)	(27%)

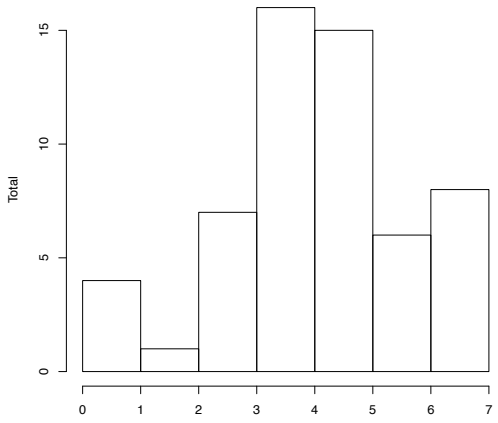

Fig. 5. Selected quality class Q_x (x-axis) in Oulu trial

Next, the campaign-wide mean expenditures and mean (selected) qualities including all the measurements (t_1, t_2, t_3) done by the users of the *Group I* and *Control* group were calculated. The mean expenditures of *Group I* and *Control*

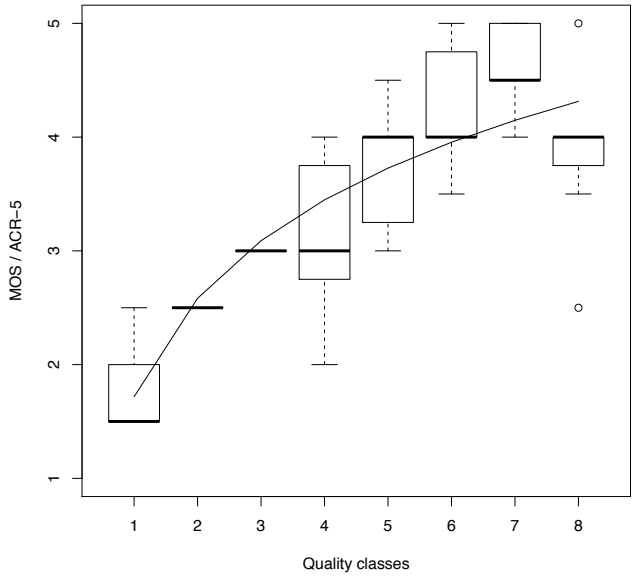

Fig. 6. Box plot of MOS ACR-5 ratings across all tariffs with logarithmic fit (*Oulu*).

groups were €2.12 and €1.65, respectively, while mean qualities were 5.97 and 4.93, respectively. Both differences were tested with t-tests and found significant on alpha level 0.05 (p-value of 0.03 for expenditure and 0.02 for quality). Similar differences can be found in Vienna trial outcome, but less significant (p-value of 0.14 for both expenditure and quality). Also, when comparing t_2 expenditure of Control group is lower than expenditure Group I (€1.52 vs €2.06) with close to significance p-value (0.11). It has to be noted that with small sample size even a single user can affect the result.

Both of these observations contribute to disproving Hypothesis 2, while we still cannot claim it to not hold.

Again focusing on t_2 and t_3 of the control group (c.f. Table V for Vienna results), the subgroups with tariff sequences $B \rightarrow A$ and $B \rightarrow C$ are compared. As shown in Tab. VIII, the results contradict those of the Vienna trial. In the Oulu trial the price drop has triggered redecision (or the users are not making an active decision, but follow an earlier price decision), while the price increases are hardly felt. Both trials show (weak) signal of price history affecting the price decision thus partially disproving Hypothesis 2.

Similarly to what happened in the Vienna trial, almost all users liked the content they chose (91 %). The subjects reported slightly higher "own money" feeling with 3.2 (ACR-5) on average (median: 3.5; "Fairly") than in Vienna trial.

V. Analysis

A. Regional Differences in Expenditures

As described in Section IV, the realised expenditures of Oulu campaign were higher than ones in Vienna. It is possible that some cultural or socio-economic factor(s) affect the purchasing behaviour of participants. On the other hand, the difference could be explained by the altered quality levels between the campaigns.

TABLE VIII
Spending as percent of p_{max} in t_2 and t_3 per control subgroup in Oulu trial.

	t_2	t_3
	B: $p_{max} = 3$	**A:** $p_{max} = 2$
Mean	57%	79%
Median	64%	79%
	B: $p_{max} = 3$	**C:** $p_{max} = 4$
Mean	46%	43%
Median	57%	57%

The observed aggregated MOS of all measurements in both Vienna and Oulu campaign was 3.8, whereas the average selected quality class in Vienna was 4.5 and in Oulu 5.5. Also, it can be observed from Figure 4 and Figure 6, that the average QoE of 3.8 is reached at Q_4 in Vienna and at Q_5 in the Oulu trial. Therefore we can conclude that on the average, participants in Oulu chose quality level one step higher compared to the Vienna trial. Furthermore, the average spending in Vienna was €1.48 and in Oulu €1.90, their difference being €0.42. The average price increment in both campaigns is €0.43 which is very close to observed average spending difference. Also, the average bitrate in Vienna trial (4080 kbits/s) is 19 % larger than in Oulu trial (3422 kbits/s). The difference is very close to difference in WTP, 22 %. It can be concluded that the findings support the Hypothesis 4: the online video service market can be regarded to be of global nature where regional limitations may be marginal and relative to the cultural diversity (applicable at least between countries within economic and cultural proximity).

B. Market Entrance and General Pricing

In the earlier work [1] it was observed that the price increases caused active decisions to be triggered (typically leading to lower normalised expenditure), while price decreases did not trigger similar re-evaluations. The similar (yet weak) effects could be observed in Vienna trial, that suggest that aggressive entry pricing (discounts) may not be suitable for all markets.

On the other hand, in Oulu trial the effects seem to indicate the opposite (admittedly, weakly). There the subjects were more willing to tolerate the price increases, once they had originally made quality selection in the t_1/t_2. This implies that for some customer segments low entry price strategy (e.g. free/discounted first month) can be a viable option.

In the Oulu trial, when the prices were dropped for part of the *Control* group (tariff B in t_2 and tariff A in t_3 as shown in Table VIII) the normalised expenditure seems to increase. This could indicate that the participants had made the initial

WTP decision already before and they were maintaining the expenditure level even after the price drop. Such customers could potentially allow introducing discount campaigns without losing much revenue while attracting new users (e.g. the existing users could get a quality upgrade for the same price).

Regarding the different set of available qualities and resulting WTP, we can observe that adding more low level quality steps has resulted in higher spending. One way to interpret this information, is to conclude that participants "had to" spend more to acquire the adequate quality (which subjects of both campaign seem to agree on). But in the end, customers spent more for the same quality in the latter trial. This highlights the importance of understanding the real willingness to pay for any offering and optimising the pricing accordingly, so that all the potential revenue gets harvested.

C. Empirical Testability

Rather high variability in the results imply that the WTP testing differs drastically from traditional QoE testing. We can identify a few factors and recommendations that may have an effect on the WTP testability:

Active decision making: WTP study typically includes an active decision component (simulating the real-life purchasing event). Unlike in the act of perceiving a stimulus (e.g. watching a video), making a buying decision requires internal evaluation considering, for example, motives, the context and the potential value of the available object (better quality in this study). Presence of such evaluation makes the cognitive processing in WTP test very different than in typical QoE test setting.

Motivation heterogeneity: Varying motivations between subjects is likely to have an effect on QoE assessments as well as WTP assessments. Additionally, in a WTP study the motives of subjects may affect also the course of the test (via active decisions), which may not happen in more passive QoE tests.

Freedom: The consistency of the laboratory test results may benefit from the high level of control. However, for a WTP to be realistic people should feel free to do the buying decisions.

Perceived gain or loss: Subjective tests do not normally have a component of perceived loss or gain, but there is a static bilateral relationship between test conductor and test participant (contribution vs. reward). However, to be realistic WTP test, the *realistic* and *strong enough* gain-loss causality must be present (e.g. better quality, smaller reward). The subjects need to feel like spending their own money facing the "pain" component of purchasing event. This could be achieved e.g. by increasing the rewards and paid prices (if possible), or using innovative rewarding schemes (e.g. using chocolate or something concrete as a currency). Alternatively, if the test design allows, the "pain" component could be for example extra (and boring) task to be done.

Difficult parameterisation: Due to the high heterogeneity of motives, content preferences, and customer segments the parameterisation of this kind of campaign is generally difficult. Only when subjects are set in a critical situation where they have to actively manage the tradeoffs between quality and price, the exploration of motives or market entrance pricing effects is possible. Otherwise only the higher-level aggregate data can be obtained that gives a rough indication on the available demand and the associated WTP. In the Vienna trial, the high noise of the content appeared to be problematic.

The comparison of the Oulu and Vienna results shows that more inadequate low-quality offers, may lead to a higher revenue (while the quality choice may be similar). Partially this could be explained by the fact that in the Oulu trial, the price (an intentional test bias) may have played a too dominant role, i.e., the quality considerations have been secondary. This could explain the high relative adaptation under price losses and highlights the need to carefully moderate the required "price bias".

Market scenario: Having the participants in the right mindset, by the creation of a realistic market scenario within the trial, despite the limitations of empirical trials in general, is crucial to obtain the required data.

Assessment methodology: The WTP studies may also benefit from simplifying the assesment tasks. For example, binary choice offers (do you want to buy this quality level for this price?) may lead to different results than scenarios offering dozens of quality classes — see [10].

Content consistency: In case a test includes a variety of different contents (like in the described work), it is necessary to harmonize them regarding the studied properties. For example, the observed quality levels that guide the purchasing decisions (as in our test during the Quality Selection Phase), should be consistent across the contents and provide similar trade-offs to be considered by users.

VI. Conclusions

The results have shown that null Hypothesis 1 has to be rejected. Substantial WTP was witnessed in both trials. Hypothesis 2 can be weakly rejected (due to lack of significance) as the historic pricing has triggered (at least weak) effects on purchasing behaviors. Interestingly, the effect of price changes varied between the trials. In Vienna trial, the price increase caused normalised expenditure to drop, while in Oulu trial it was price decrease that triggered more significant effect, causing normalised expenditure to rise. Hypothesis 3 is rejected as there are different distinguishable customer segments (in Oulu trial), although not as clearly as in previous studies. Finally, the Hypothesis 4 is considered proved as the average WTP was on the same level in both trials (after compensating the effect caused by different quality levels between the campaigns).

Regarding the pricing aspects of video services, the results indicate that for some video streaming services the same pricing scheme can be applied successfully to different regional markets of the same geopolitical area. On a global market level, where cultures are targeted that do not moderately resemble each other culturally and economically, moderate differences may still be observed, which cannot be answered quantitively from the conducted trials.

We also repeat the recommendation of earlier work, that companies need to be cautious about extremely low teaser discounts as increasing the prices later on can be challenging. On the other hand, for some customer segments/cultures the service/quality level lock-in may prove to be strong enough to allow later tariff increases (as indicated by Oulu results). Finally, the results imply that clever pricing and packaging of the same product (additional lower quality levels in this work) can potentially increase the profits in some cases.

The trials have shown the challenging nature of conducting WTP trials (compared to traditional QoE testing). Some challenges were identified and recommendations are given in Section V-C. The future retestings shall take these further recommendations into account in test design.

VII. Acknowledgments

Martín Varela's and Toni Mäki's work was partially funded by Tekes, the Finnish agency for research innovation, in the context of the CELTIC+ project NOTTS. The research leading to these results has partially also received funding from the European Community's Seventh Framework Programme for the PRECIOUS project under grant agreement no. 611366.

References

[1] P. Zwickl, A. Sackl, and P. Reichl, "Market Entrance, User Interaction and Willingness-to-Pay: Exploring Fundamentals of QoE-based Charging for VoD Services," in *Proc. of the IEEE Globecom'13*, 2013, pp. 1310–1316.

[2] N. Staelens, P. Coppens, N. Van Kets, G. Van Wallendaef, W. Van den Broeck, J. De Cock, and F. De Turek, "On the impact of video stalling and video quality in the case of camera switching during adaptive streaming of sports content," in *Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, 2015, pp. 1–6.

[3] C. Keimel, A. Redl, and K. Diepold, "The TUM High Definition Video Datasets," in *Proc. of the Fourth International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, 2012, pp. 97–102.

[4] International Telecommunication Union, "Subjective video quality assessment methods for multimedia applications," *ITU-T Recommendation P.910*, April 2008.

[5] *ITU-R BT.710-4 Subjective assessment methods for image quality in high-definition television*, ITU-R BT.710-4, ITU, 1998.

[6] F. Agboma and A. Liotta, "Qoe-aware qos management," in *Proc. of the Sixth International Conference on Advances in Mobile Computing and Multimedia*. ACM, 2008, pp. 111–116.

[7] K. Ivešić, L. Skorin-Kapov, and M. Matijašević, "Cross-layer QoE-driven Admission Control and Resource Allocation for Adaptive Multimedia Services in LTE," *Journal of Network and Computer Applications*, 2014.

[8] M. Varela, P. Zwickl, P. Reichl, M. Xie, and H. Schulzrinne, "Experience Level Agreements (ELA): The Challenges of Selling QoE to the User," in *Proc. of the ICC 2015 Workshops*. IEEE, 2015, pp. 1741–1746.

[9] P. Nelson, "Information and Consumer Behavior," *Journal of Political Economy*, vol. 78, no. 2, pp. 311–329, 1970.

[10] P. Zwickl, P. Reichl, L. Skorin-Kapov, O. Dobrijevic, and A. Sackl, "On the Approximation of ISP and User Utilities from Quality of Experience ," in *Proc. of the Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, 2015.

[11] FP5 Project M3I, IST–1999–11429, "Deliverable 15/2 – M3I user experiment results," 2002.

[12] P. Reichl, P. Maillé, P. Zwickl, and A. Sackl, "A Fixed-Point Model for QoE-based Charging," in *Proc. of the ACM SIGCOMM Workshop on Future human-Centric Multimedia Networking (FhMN)*, 2013, pp. 33–38.

[13] A. Sackl, S. Egger, P. Zwickl, and P. Reichl, "The QoE Alchemy: Turning Quality into Money. Experiences with a Refined Methodology for the Evaluation of Willingness-to-pay for Service Quality," in *Proc. of the Fourth International QoMEX Workshop*. IEEE, 2012, pp. 170–175.

[14] A. Sackl, P. Zwickl, S. Egger, and P. Reichl, "The Role of Cognitive Dissonance for QoE Evaluation of Multimedia Services," in *Proc. of the 2012 IEEE Globecom Workshops*. IEEE, 2012, pp. 1352–1356.

[15] *ISO/IEC 23009-1:2014: Information Technology – Dynamic Adaptive Streaming over HTTP (DASH) – Part 1: Media Presentation Description and Segment Formats*, ISO/IEC 23009-1:2014, International Standards Organization (ISO), 2012.

[16] *H.264: Advanced video coding for generic audiovisual services*, Recommendation H.264 (02/14) (twinned), International Telecommunication Union (ITU), 2014.

[17] *H.265: High efficiency video coding*, ITU-T H.265 (V3) (04/2015 (twinned), International Telecommunication Union (ITU), 2015.

[18] *ISO/IEC 23008-2:2015: Information technology – High efficiency coding and media delivery in heterogeneous environments – Part 2: High efficiency video coding*, ISO/IEC 23008-2:2015 (twinned), International Organization for Standardization, 2012.

[19] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.