Fix Your Netflix Experiment

Lucjan Janowski, Margaret H. Pinson, Dominika Wanat, Kamil Koniuch, Katrien De Moor, Mark D. Gross

***Abstract*—This paper examines the hypothesis that asking people about video quality changes their behavior. We conducted an Absolute Category Rating (ACR) scale test as a baseline. We also conducted an experiment in which subjects chose a movie to watch and pressed a button whenever the quality disturbed their watching experience. This experiment design could lead to improved understanding of bitrate ladders. The action-focused experiment, whereby most subjects tolerated only very high-quality video, revealed an extremely large diversity of button-pressing behaviors. Our analyses indicate that each subject has their own unique lower threshold for the video quality that they tolerate. We observed two subgroups of subjects, one demanding and the other relaxed. The button pressing threshold of the demanding subgroup maps to between good and excellent on the ACR scale; that of the relaxed subgroup maps to between fair and good. An area for future work is an algorithm that differentiates between demanding and relaxed subjects.**

***Index Terms*—QoE, quality of experience, subjective test, user experience, video quality**

I. Introduction

This is the second in a series of papers investigating the context and factors that influence quality of experience (QoE) when people use video services [1]. The goal is to develop experiment designs that more accurately measure QoE. This paper presents a video quality experiment in which subjects were not asked about quality, to examine the hypothesis that asking people about quality changes rating behavior.

ITU-T Rec. P.910 [2] and ITU-R Rec. BT.500 [3] describe best practices for conventional subjective tests that assess video quality. With the most popular method, Absolute Category Rating (ACR), subjects watch a short video and then explicitly rate the quality. These ratings are averaged across a pool of subjects to calculate a mean opinion score (MOS). These well established methods inspire trust and lead us to assume, without proof, that they accurately measure all aspects of QoE.

From a comparison of six labs in 2012, conducted by the Video Quality Experts Group (VQEG), we know that these are relative MOSs, not absolute MOSs [4]. We can reach strong conclusions about relationships among videos, like "system A is significantly better than system B," but when the same videos are rated in multiple labs, statistical tests may show significant differences. We cannot make comparisons to constant MOS thresholds, because the top and bottom of the scale is unpinned—the effective meanings of excellent and bad shifts to accommodate the range of quality presented, the set of subjects involved, test environment, user expectations, etc.

From our modeling of subject ratings in 2015 [5], we learned that subjective ratings is a random process that is influenced by three random variables: subject bias, subject inaccuracy, and stimulus inaccuracy. The observed distributions for these three variables span about +/- 25% of the ACR rating scale. These random variables explain apparent inconsistencies within a single subject's data and probably cause much of the lab-to-lab differences seen in datasets scored at multiple labs. The newest statistical analysis method in ITU-T Rec. P.910 builds on this theory to weight subject ratings by their reliability when calculating MOSs [6].

Our study on the accuracy of subjective tests in 2023 [7] proves that conventional subjective tests are highly precise and repeatable. We used a confusion matrix to compare conclusions reached by 88 lab-to-lab comparisons, 22 method-to-method comparisons, and 12 comparisons between expert and naïve subjects. We used the disagree incidence rate to identify lab-to-lab differences (i.e., the likelihood that significantly different stimulus pairs have opposing rank order from the two labs). We concluded that disagree incidence rates above 0.31% are unusual enough to warrant investigation and disagree incidence rates above 1.0% indicate differences in method, test environment, test implementation, or subject demographics.

Observations at VQEG meetings indicate that companies often treat MOSs as absolute and rely upon MOS thresholds. An MOS threshold of "good quality or better" is generally recommended for purposes like the minimum expected quality for paid video streaming services or the minimum allowed quality of source video sequence in video quality metric validation tests. Based on our ad-hoc observations, VQEG subject matter experts are aware of all three studies and acknowledge the validity of this prior work.

So, which assessment is more accurate? Is the relative nature of MOSs a minor factor, such that it is only of interest for statistical tests performed on lab-to-lab comparisons? In that case, the general guideline of "good or better" would be reasonable within the context of a specific scenario, such as a paid service streaming 1080p 24fps video to a laptop. Does the action of asking people to assess quality—and the scale presented—change how people think about quality? Perhaps conventional subjective tests focus attention only on perceptual drops in quality (relative differences), while diminishing or eliminating people's opinions on whether the quality is objectionable (e.g., whether they care enough to act). In that case, conventional subjective tests and MOS thresholds cannot accurately predict people's actions in the real world, such as deciding whether to change service providers.

In order to address the above described question, we have

L. Janowski, D. Wanat, and K. Koniuch are with AGH University of Krakow. M. H. Pinson is with the National Telecommunications and Information Administration's Institute for Telecommunication Sciences (NTIA/ITS). K. De Moor is with Norwegian University of Science and Technology - NTNU. M. D. Gross is with University of Colorado Boulder.

The research leading to these results has received funding from the Norwegian Financial Mechanism 2014–2021 under project 2019/34/H/ST6/00599. Cooperation with Margaret H. Pinson and Mark D. Gross was possible thanks to project Bekker BPN/BEK/2023/1/00228.

to link user behavior and quality. Our initial claim was:

- If we measure behaviors but do not ask about quality, our conclusions will differ from the results of conventional video quality experiments.

To explore this claim, we will implement a modified experiment protocol that avoids asking people about quality directly, due to concerns that such questions could influence people's opinions. During the viewing, the video quality will fluctuate. If at any point the quality interferes with the subject's experience, they are instructed to stand up and approach the TV to push a button that resets the simulated service to the highest possible level of quality. The subject will then participate in a conventional subjective test. We will use the Video Multi-method Assessment Fusion (VMAF) metric [8] as a bridge to compare the button-pressing behavior with MOSs.

This experiment is designed to measure the subject's willingness to act in the lab. Our theory is that such actions taken in the lab may be more indicative of actions in real life than opinions expressed. This experiment design has reduced internal validity, as compared to a conventional subjective test, and it increases mundane realism with the hope that external validity will also improve [9]. Although this introduction focuses on our work, these efforts were only possible due to contributions from many other researchers and insightful discussions during VQEG meetings.

The main contributions of this paper are:

- A new experiment design where people perform a physical action instead of assessing quality on a written scale
- Methods to analyze the data obtained from this experiment
- Methods to compare action-based experiment data with conventional ACR data
- An improved method to estimate the adaptive bitrate streaming (ABS) ladder, based on user behaviors and quality tolerances

II. State of the Art

Our experiment compares the conventional experiment design with a design introducing actions by subjects. Therefore, this section is divided into two parts. The first analyzes articles written for consumers and scientific publications that examine consumer behaviors in real world. The second part lists related research papers that describe QoE experiments. We need the first part to clearly describe why we propose an experiment that does not pose questions about or direct subjects to consider quality.

A. Talking to Consumers

Conventional subjective tests must display varying levels of quality, or the task gets too frustrating. This is unrealistic. Real video streaming services present a stable level of video quality. Quality fluctuations are usually caused by the rare burst of network errors (e.g., packet loss or rebuffering).

Robitza *et al.* [10] installed an app on subjects' phones to monitor subjects' use of Amazon Prime Video™, Netflix®, and YouTube™ services[1]. The authors collected playback parameters like usage rates, abandonment rate, and loading times for 447,489 viewing events and more than 2,000 viewers. The network parameters were fed into a bit-stream metric, to estimate MOSs. The authors, when they presented this study to VQEG, observed that their subjects typically did not encounter dynamically changing quality. This phenomenon can be observed in the quality distribution histograms (Fig. 7 of [10]), where most data lies between good and excellent, and in the observation that stalling was responsible for most of the drops in quality.

These results are in line with a quality study of different applications, which shows that the applications used were good or excellent [11]. This is a logical consequence of testing in a real environment. The applications that are actually used have to be at least of good quality. Otherwise, the application is not used and we do not see them in the real use cases.

Conversely, people will tolerate very low quality for specific purposes. Examples include very old films, videoconferencing during social isolation, and a sports fan who can only gain access to a particular event via a very low quality video stream.

We interviewed more than 100 first responders about problems with image and video quality [12], and how they spoke about quality differed greatly from conventional subjective tests. People naturally expanded the scope of our discussions to include problems with camera systems in general. These first responders identified a large variety of problems, such as:

- Form factors (e.g., dirty cellphone lenses and bodycams pointing away from the action)
- Obscurants (e.g., surveillance system cameras blocked by spider webs, growing vegetation, dust storms, rain, and falling snow)
- Environmental conditions that hinder the camera's performance (e.g., low light conditions and different lighting conditions in the same scene)
- Process flow (e.g., incompatible video file formats, data storage, incorrect time codes, and inaccurate geotagging)
- Camera capture system (e.g., poor telepresence, mismatches with human perception, and lack of accurate colors)

First responders' opinions of video quality depended on the task being performed. The detectives and forensic video analysts were adamant in stating that a threshold for usability could not be established, because any video evidence was useful. Color accuracy matters more for first responder use cases than for entertainment use cases, and we were eventually able to detect the impact of this on subject ratings [13]. However, in casual discussions, first responders rarely talked about pixel quality specifically.

To assess how novices think about video quality, we searched the internet for consumer articles that assessed the quality of video streaming services and were published within

[1]Certain commercial equipment, materials, and/or programs are identified in this report to specify adequately the experimental procedure. In no case does such identification imply recommendation or endorsement by the National Telecommunications and Information Administration or AGH University of Krakow, nor does it imply that the program or equipment identified is necessarily the best available for this application.

the last five years. We limited our analysis to nine articles from well-known consumer organizations. We will examine these articles individually and then note trends.

Only three articles mention video quality. *Compira Labs* [14] recommends five KPIs: video startup time, rebuffering, bitrate, resolution, and latency. *Wired* [15] considers two of these KPIs (resolution and bandwidth) plus user contracts, and ad-supported vs ad-free content. *CableTV* [16] considers two KPIs (resolution and rebuffering) plus cost, user contracts, variety of channels, and the available content.

The other six articles do not mention video quality, and therefore we interpret this to mean that video quality does not matter. *PCMag* [17] assesses the best video streaming services by comparing variety of content, ad-supported vs ad-free content, cost, and user interface. *Forbes* [18] compares streaming behaviors and preferences by assessing time spent streaming, likelihood of having subscriptions, cancellation rates, cost, ad-supported vs ad-free content, revenue, and popularity. The *Consumer Reports* guide on streaming services [19] focuses on cost, content library, and ad-supported vs ad-free content. The *Nielsen* analysis of video streaming consumption [20] mentions time spent streaming, growth of the industry, the large number of services, the increase in content, and types of services. The *PricewaterhouseCoopers* (*PwC*) video streaming survey of 1,000 consumers [21] investigates many topics—emotional response, variety of content, viewing devices, subscriptions, user interfaces, personalized adds, comparisons with cable TV, reason for leaving a show early, strategies for retaining customers, etc.—but not video quality. The *Whip Media* analysis of streaming video services [22] focuses on user satisfaction, likelihood to keep services, favorite services, quality of programming (content satisfaction), variety of content, content suggested by the service, cost, customer churn, reasons for canceling a service, and user experience.

None of these nine consumer articles acknowledges differences among the services' video encoders, transmission strategies, and bitrates. From private communications, we know that the quality of video encoders differs. The impact of transmission strategy on quality can be seen in the research of Yang *et al.* [23], who propose an adaptive video streaming scheme for intercity railways, and Minh *et al.* [24], who propose strategies for adaptive streaming to benefit from new HTTP features. Bandwidth is assumed to be a variable controlled by the consumer (e.g., "Just about every streaming service is smart enough to adjust the video quality in response to the available bandwidth on the Wi-Fi (or cellular) network that you're connected to" [15]). None of the articles acknowledges the service providers' need to balance multiple factors when designing their content delivery network (CDN), including per-title bitrates, and constraints on multiple storage spaces around the world.

These consumer articles do not reflect the interest that streaming service providers express about video quality, as reflected by their participation in VQEG. This discrepancy is not an issue of perception; conventional experiments show that subjects are able to distinguish quality when asked to focus on this task [7]. Therefore, we have no truth data and an unclear relationship between video quality and people's actions in the real world.

We also observe anecdotal evidence that quality matters. Significant effort has been dedicated to improving the quality of old movies and photographs, as demonstrated by numerous publications on quality enhancement algorithms [25], [26]. Another indication is the popularity of high-quality short videos on Instagram™® and YouTube [27], which are typically of good or excellent visual quality [28]. The equipment, lighting, and production techniques used by professional content creators on social media suggest that video quality is a priority. Although their gear is generally less expensive than broadcasters' cameras, it still surpasses what is achievable with a basic smartphone. Finally, even children tend to notice quality differences, often commenting that low-quality content "looks like a photo from the 1980s." These observations indicate that visual quality is perceptible and valued—though its impact may be less immediate or less critical when comparing services that are already technically similar.

B. Related Experiments

Related experiments can be divided into two different categories. The first concerns analyzing real data to obtain the quality thresholds from observed behaviors. The second concerns subjective experiments that pose no questions about either quality or preferences. We will start with analyses of real data.

Robitza *et al.* [10] collected and analyzed data from 400,000 video playback events across three major video streaming services. Unfortunately, their analyses were not able to reach strong conclusions on the influence of video quality on user engagement. This is left for a future study, but we did not find such a publication. Either the relationship between quality and engagement was too difficult to extract from the real world data, or perhaps the infrequency of such events was too low (e.g., because the study was limited to major streaming services).

Huan *et al.* [29] performed a giant study of ISP data, which studied the migration of 3.8 million users among content providers (CP). They concluded that poor service quality rarely causes users to migrate between CPs (meaning stalling or a requested video not playing). Instead, in nearly 60% of migrations, the user played one of the most popular videos. The video quality (pixel quality) is not mentioned, likely because that information is not currently available in the ISP data.

Leszczuk *et al.* [30] provides information on the methods provided by ITU-T Rec. P.912 [31], Subjective video quality assessment methods for recognition tasks. The idea is to design videos that can be used to perform a task that has a correct answer (e.g., reading a license plate). The goal of the experiment is to quantify video system characteristics that permit the task to be performed (e.g., minimum resolution and bitrate). The advantage is that the method is deterministic; there is a right and wrong answer. The disadvantage is that it is difficult to collect or create source videos that demonstrate all aspects of the tasks (e.g., object size, scenery, amount of motion in the scene, and camera impairments).

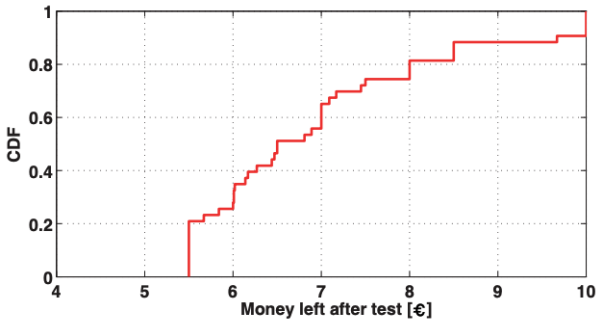


Fig. 1: Figure, reproduced with permission from Sackl *et al.* [32], showing distribution of spending for the quality improvement. We see that 20% testers spend all the money, and the other 20% spend a maximum of 15% of the available budget.

Sackl *et al.* [32] use monetary incentives as a substitute for subject ratings, to explore the impact of network quality on video on demand (VoD). Each subject chose three videos with a total duration of one hour. At the beginning of each video, the subject explored four network quality of service (QoS) levels to understand the impact of each QoS level on video quality. The subject had a small pool of funds that could be spent on increased QoS, and thus video quality. Any unspent funds were theirs to keep. As shown in their Fig. 5 reproduced here by permission of the authors (see Fig. 1), subject responses spanned the full range from investing none of their own money (10% had €10 remaining) to paying the maximum amount allowed for enhanced video quality (20% had €5.5 remaining). We will refer to this study later, in our data analysis.

The FYN experiment design is, in some ways, similar to the very popular Just Noticeable Difference (JND) method [33]. JND starts by using a single source video to create a large pool of stimuli (e.g., different compression bitrates). Subjects compare the source video (called JND_0) with stimuli from the pool until they find a stimuli where 75% of subjects say JND_0 is "better" than this stimuli (i.e., the quality difference is noticeable to 50% of subjects and the other 50% of subjects answer randomly). This is JND_1. We repeat this process by comparing JND_1 with stimuli from the pool until we find JND_2, etc. The final JND dataset contains a series of videos, named JND_0, JND_1, JND_2, ..., JND_n [34].

FYN is similar to JND in that both experiments search for a reaction and a threshold. Therefore, the analytical solutions are similar to those considered in [35]. Also, generalization of JND beyond just JND experiments [36] is an interesting step which we will explore in our analyses. The similarity of the data obtained (probability as a function of quality) and the analytical method (looking for the correct fitting curve) are the only true similarities. In the FYN experiment, we do not ask if subjects see a difference but if they care enough to stand up and act, despite interrupting an interesting movie.

Miao *et al.* [37] and Brunnström et al. [38] provide examples where video quality is assessed without asking people about quality. In each, the subjects perform a task, and the task performance is the quality measure: backing up with a digital rear view camera [37] and operating a VR system to load a logging truck [38]. Our idea of conducting an experiment without asking about quality is often used by User Experience (UX).

III. Description of Experiments

We conducted two different experiments. The first, Fix Your Netflix (FYN), was an experimental method in which subjects acted instead of rating quality. The second, Conventional VMAF Test (CVT), was a conventional subjective test.

In this paper, we will always begin by describing CVT first, because it provides a known basis for comparison. However, subjects performed FYN (March 7 to June 17, 2023) before they performed CVT (October 9 to 30, 2023).

For the CVT experiment, we invited the same subjects who participated in the FYN experiment so that we would able to make within-group comparisons. Since we wanted to have a diverse pool of subjects, we advertised the experiment in two different places: the Facebook platform and a senior association. The first resulted mostly in young people; the second, mostly in people over 50. The age distribution is presented in Fig 2. The blue color indicates all participants who passed the data cleaning procedure described in Section IV-A.

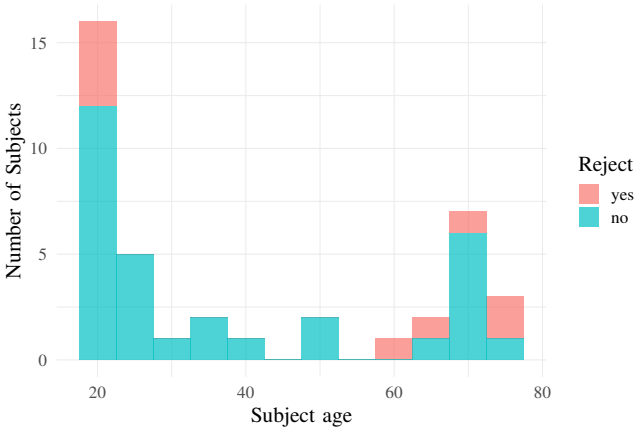


Fig. 2: Age distribution for all participants marking who passed the data cleaning procedure (see Section IV-A)

A. VMAF as Proxy for MOS

Our experiment design has two very different experimental procedures. FYN is a reaction to quality. CVT is a quality judgment. We cannot use the same sequences in both procedures, since reacting to quality requires engagement, and engagement requires personal content selection. On the other hand, the conventional experiment assumes that all participants rate the same sequences. In order to compare results from both experiments, we need to know for both experiments what quality was watched before the rating or action, respectively. We used the Video Multi-method Assessment Fusion (VMAF) metric, which is publicly available [8]. More details about why we selected VMAF can be found in our forthcoming publication [9].

B. Conventional VMAF Test (CVT)

CVT is a conventional experiment with repeated scenes. The source sequences (SRC) are 5-second stimuli extracted from movies on the "Netflix Open Content" website [39]. We used HEVC/H.265 to encode a large pool of processed video sequences (PVS) using a full matrix of constant rate factor (CRF) levels from 20 to 35 and resolutions from 4K to 480×270. We calculated VMAF for each SRC and PVS, and then selected stimuli that are close to 10 evenly spaced VMAF scores (i.e., 100, 89, 78, ... 0). This allowed the CVT to have a uniform quality distribution. Note that VMAF=100 for all SRC, because VMAF is a full reference (FR) metric. The CVT dataset contains 171 stimuli: 12 SRC with 11 PVSs, 3 SRC with 10 PVSs, and 1 SRC with 9 PVSs.

The CVT subjective experiment was conducted in a dedicated laboratory with light levels set according to the ITU-T Rec. BT.500 specification. The stimuli were played on a 55-inch 4K television at a distance of three picture heights (3H). The experiment was controlled by the AVRateNG software [40], [41]. Subjects rated the stimuli in a random order on the 5-point ACR scale. Of the 40 subjects who participated in the FYN experiment, 35 returned for the CVT experiment. The CVT dataset has 35 ACR responses for each of the 171 stimuli.

C. Fix Your Netflix (FYN)

The FYN experiment focuses on reacting to quality instead of rating quality. Subjects selected any movie from the Netflix platform. If a subject did not know what to watch (17 subjects out of 40 subjects), we proposed three pre-selected movies that had good reviews. After selecting a movie, each subject sat on a sofa positioned at a distance of 3H from the 55-inch television (approximately 1.5 meters).

The recorded instructions (translated from Polish) were as follows: "You can set up the seating position as you want, even shifting some furniture if you want. While watching, the quality can drop. The moment the quality disturbs your watching experience, approach the television and press the button that is on the left side of the television. After you press the button, the quality will be changed to the best available. Do not move the button; it should stay at the same position." The button pressing action also paused the playback briefly and rewound the video for 1 to 2 seconds. None of the subjects shifted the furniture.

The system[2] decreases quality every 90 seconds (± jitter), so that subjects could predict when the quality will drop. The quality drop was implemented by selecting the next lower bitrate from the Netflix service. Each time the button was pressed, the quality was raised to the maximum available bitrate. After the movie ended, the subject filled out a questionnaire.

We record all the information available from the diagnostic screen of the Netflix service (e.g., link to the movie, bitrate played, resolution played, VMAF rating, and buffering information). The system estimates VMAF as a constant for the entire movie at each bitrate (i.e., ignoring fluctuations over time). However, since the Netflix content processing algorithm [42] targets specific VMAF levels, these estimates are reliable. We extracted and recorded the names of movies watched in April 2025, when we noticed that some of the movies links were broken. The FYN dataset contains the VMAF values as a time series, and the times when subjects pressed the button.

[2]https://github.com/TUFIQoE/FixYourNetflix-TUFIQoE-2022/blob/main/README.md

IV. Data Analysis: Data Preparation

This section prepares the data from our two experiments. We start by cleaning the data, and then extract parameters for each experiment.

A. Data Cleaning

Let us begin by cleaning the data. We want a within-group comparison[3] between the FYN and CVT experiments, so we must discard data from the FYN subjects who did not return for the CVT experiment (subjects 8, 17, 25, 33, and 37).

Using ITU-T Rec. P.910 Annex A [2], we will be screening the CVT subjects. Fig. 3 shows the Pearson correlation between each subject and the overall CVT MOS. Based on the distribution formed by the majority of CVT subjects, the appropriate threshold for this test is around 0.65 to 0.60. Therefore, we will only remove Subject 42, whose correlation is a clear outlier for CVT. Subject 42 also pressed the button erratically in the FYN experiment.

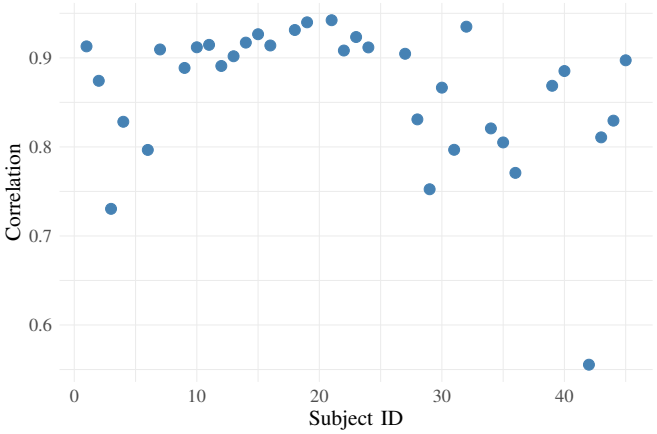


Fig. 3: Correlation between each CVT subject's ratings and the overall CVT MOS.

To screen the FYN subjects, we will perform visual analyses of the button pressing time series. The FYN data and FYN questionnaires indicate atypical button-pressing behaviors at the beginning and end of the movie (see Fig. 4). At the beginning, some subjects are not sure how the system worked and are slow to press the button. At the end, some subjects do not want to interrupt the finale or perhaps do not care about the quality of the closing credits. Therefore, we will omit all data up to and including the first button press, and we will omit all data after the last button press.

[3]A within-group comparison examines the behavior of identical subjects who participate in two different experiments, allowing us to compare behaviors in both experiments.

We will eliminate Subjects 12, 29, and 34 because they never press the button. The questionnaire only provides insights into subject 34 (who did not understand the test instructions) and Subject 12 (who claims that the quality never dropped low enough). We cannot rule out the possibility that Subjects 12 and 29 do not want to admit that they did not understand the instructions. While it would be interesting to include these extreme behaviors, their data would always be a special case with unknown accuracy.

Invalid FYN data are discarded. Subject 26 changes movies part way through FYN, because the original movie was too scary. We will discard the data from their original movie. The FYN test software failed for Subjects 5 and 20. Subject 41 was an internal test of the FYN system.

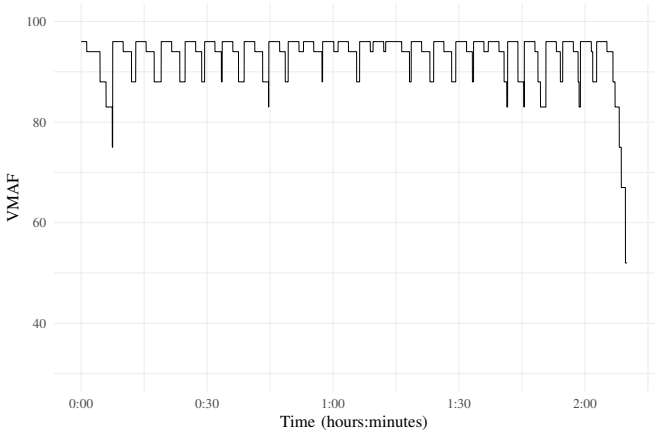


Fig. 4: Quality fluctuation for subject 40. The first and last button pressing behavior are different from the middle of the experiment. VMAF scale is synchronized with Fig. 6.

After this data cleaning, we have 31 subjects. Each subject scored 170 sequences in CVT and pressed the button between 6 and 25 times in FYN, for an average number of 10.7 button presses.

B. Extracting Data from CVT

The CVT experiment produced a Pearson correlation of 0.92 between VMAF and MOS, with a typical spread around the fit line (see Fig. 5). Since we are mostly interested in a subject's behavior, we used the "Bias-subtracted consistency-weighted MOS method" described in P.910 [2] and [6] to improve our MOS estimations. This method estimates the subject bias (Δ_i) and the subject consistency (σ_i) for each CVT subject.

C. Extracting FYN Data Trends

The FYN experiment generates a time series for each subject that shows which VMAF was seen and for how long. Fig. 6 shows an example of three very different button pressing behaviors. Subject 12 (mentioned earlier) watched a short movie and did not press the button. Subject 19 pressed the button fairly consistently, when VMAF was around 80 to 90. Many subjects' data show similar button pressing behaviors, but with different thresholds. Subject 3 watched a long movie

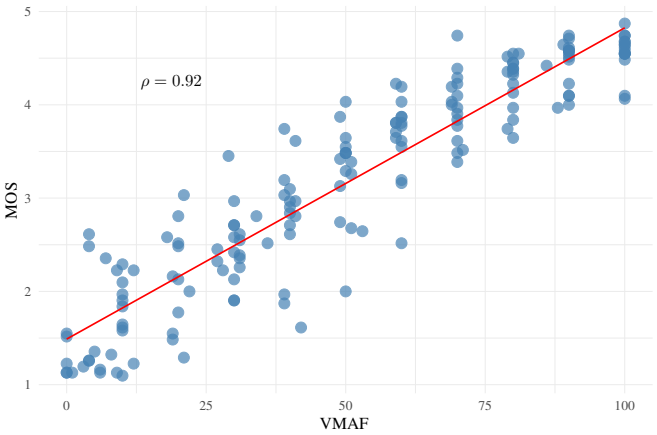


Fig. 5: Correlation between VMAF and MOS for CVT experiment.

and had a large variety of button pressing behaviors. Sometimes the button was pressed quickly ($\approx$ 95 VMAF at time 1:00); other times, the subject watched for several minutes before pressing the button ($\approx$ 35 VMAF just before 1:30).

Fig. 6 shows why we have different numbers of FYN responses for different subjects. If someone keeps the quality high, the button is pressed more often, generating more data. If someone watches a longer movie, we gain more data.

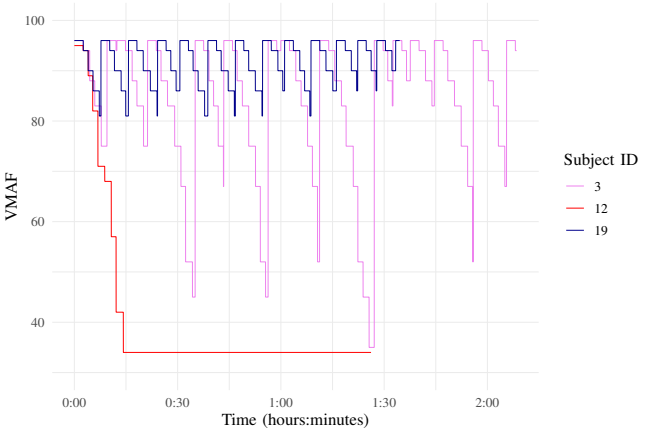


Fig. 6: Pressing process in time for three FYN subjects with different behaviors.

The FYN time series data can be used to generate a more standard representation, focusing on which level makes a subject stand up and take action. The Method of Limits is commonly used for this purpose in psychophysics experiments. The Method of Limits is an experiment paradigm that tries to detect a response threshold. In our case, this threshold is the distortion level at which the subject will typically press the button. A simple example is a study to detect the temperature sensitivity threshold as a function of age and where subjects were touched [43].

Our data are very imprecise, because the server only has around 7 to 10 VMAF levels for each movie. Nevertheless, the goal is not to show all the cases, but to estimate the sensitivity

function. We know that for perfect quality, no subject is interested in pressing the button. For quality where nothing can be seen at all, each subject should be interested in fixing it. Using these assumptions, we change the time series into a probability plot (see Fig. 7).

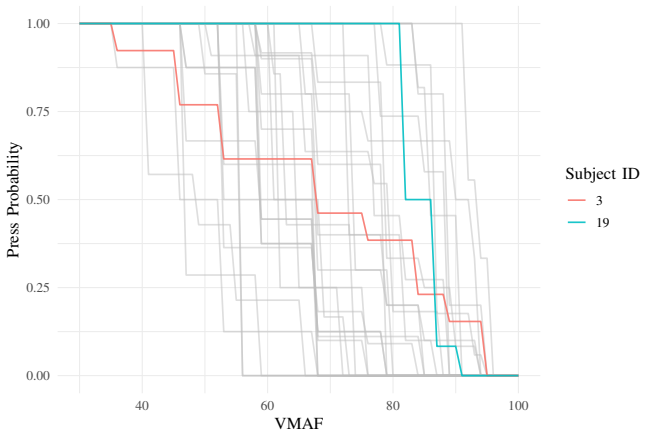


Fig. 7: Probability for each subject of pressing the button. Subjects 3 and 19 are highlighted (see Fig. 6).

We want to find the shape of each subject's pressing probability function. As per best practices in psychometry, we will fit the Generalized Linear Model (GLZ) to our data. This yields a function that predicts the probability of a particular subject pressing the button, as a function of VMAF. For each subject, we fit the psychometric function using R package `quickpsy` [44]. Examples are presented in Fig. 8.

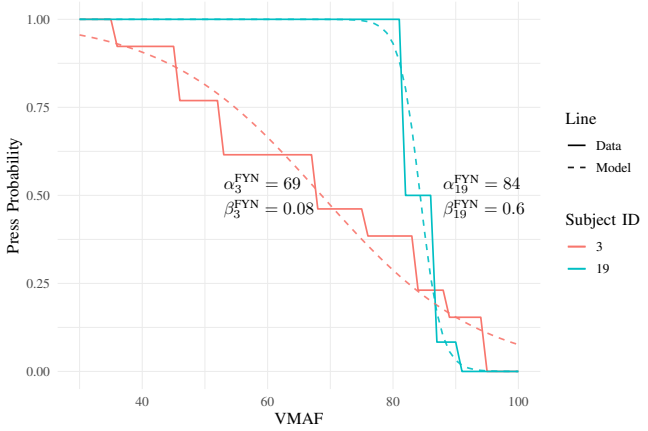


Fig. 8: Button-pressing probability functions for subjects 3 and 19, estimated using GLZ. We are interested in the model parameters: threshold (α) and slope (β).

The psychometric function, given by

$$f(v) = \frac{1}{1 + e^{-\beta_i(-v+\alpha_i)}} \tag{1}$$

allows a description of each subject's threshold (α_i), which can be understood as sensitivity. For VMAF $= \alpha_i$, a subject has a 50% chance of standing up and pressing the button. Mathematically, the threshold is a value of v (in our case, VMAF) for which the psychometric function reaches 0.5. The second parameter is slope (β_i), which can be understood as the subject's consistency. Mathematically, the slope is the value of divergence at the threshold point; so it is how fast the psychometric function changes at the threshold. Fig. 8 presents example data and models for two subjects.

The interpretation of the parameters of the psychometric function for video quality is as follows: The threshold is the level of quality for which the probability that the subject takes an action is 50%. A lower threshold indicates that a person can watch a lower-quality video without the willingness to react. The slope determines the size of the threshold confidence interval by $\sim \frac{1}{\beta_i}$ [45]. For slopes close to 0, the threshold confidence interval is wide. Such a person might have pressed the button in response to a quality level of 95 or 35, or anywhere in between, as shown for subject 3 in Fig. 6 and Fig. 8.

As always in data analyses, we should distinguish the theoretical value α_i and the estimated value, traditionally marked by $\hat{\alpha}_i$. Since we are going to estimate α_i by two different methods, we will use α_i^{FYN} if the value is estimated by FYN data. The α or α_i refer to a theoretical value. Any value marked by a dataset name means an estimate based on that specific dataset.

Fig. 9 summarizes the results. We see that the threshold is widely spread, with very close to a uniform distribution.

V. Data Analysis: Comparing Behavioral and Conventional Tests

For both FYN and CVT, the same subjects can be described by parameters that describe their sensitivity to quality (α_i and Δ_i) and the stability of their answers (β_i and σ_i). We would like to understand how those two values relate to each other. Because we have so little data, our estimates for stability are inaccurate. We will therefore focus on sensitivity measured either directly (using α_i and Δ_i) or indirectly (using VMAF).

A. Comparing Sensitivity

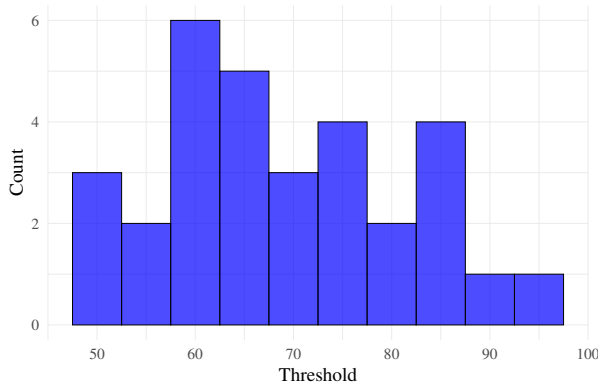
Fig. 10 shows a scatter plot between α_i and Δ_i. The Pearson correlation is -0.37 with a p-value of 0.04 (i.e., a weak but statistically significant relation). This means that Δ_i only explains around 14% of the variance in α_i.

Since the two experiments are very different, this confirms the existence of personal preferences for higher or lower quality, but the relation is rather weak. It also suggests that conclusions drawn from ACR experiments cannot be directly mapped to real world behaviors. This analysis is imperfect because Δ_i is generated by averaging opinions over all possible qualities, while α_i is based on the qualities people would like to see.

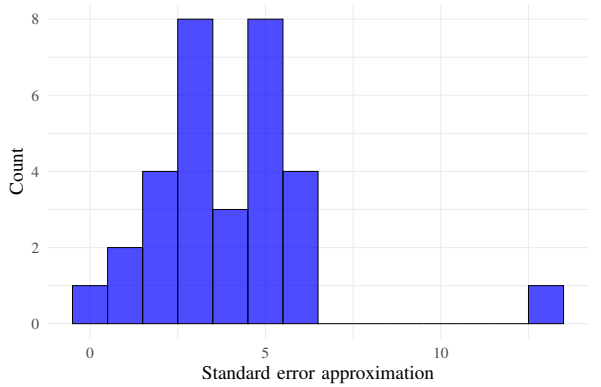
B. VMAF Distributions

Fig. 11 shows the distributions of VMAF levels presented to subjects for the FYN and CVT experiments. Subjects are sorted and color coded by FYN sensitivity to quality, α_i.

The CVT distribution is fairly uniform, and each VMAF bin has an identical distribution of colors. That is, the CVT



(a) Histogram for threshold estimation (α_i^{FYN}).



(b) Histogram for standard error approximation ($\frac{1}{\beta_i^{\text{FYN}}}$).

Fig. 9: Summary of the parameters extracted from FYN by psychometric function.

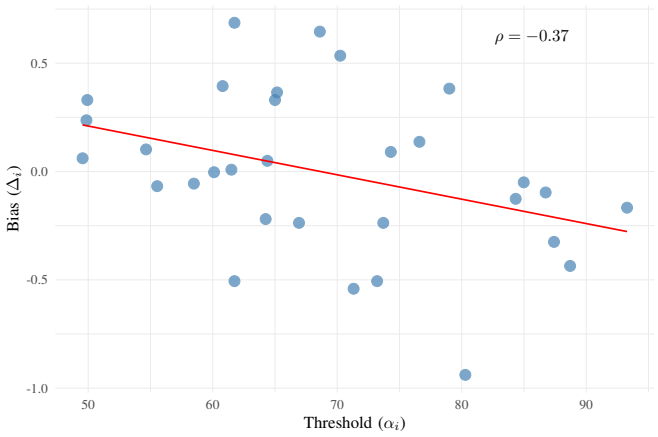
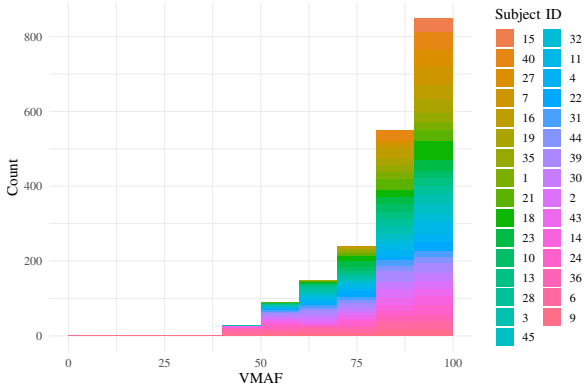


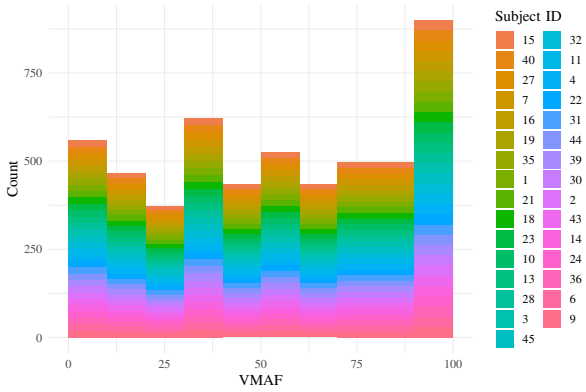
Fig. 10: Scatter plot and correlation for α_i and Δ_i obtained for FYN and CVT experiments.

stimuli span the full range of quality from VMAF 0 to 100, and all subjects saw the same range of quality.

The FYN distribution is skewed to the right: predominantly VMAF $>$ 80 and rarely VMAF $<$ 50. This is a natural consequence of the design of the experiment. To see poor quality, the subject has to watch and not react as quality drops from excellent, to good to fair to poor. Also, the proportion



(a) VMAF Distribution for FYN



(b) VMAF Distribution for CVT

Fig. 11: VMAF distribution for both experiments. The order of subject is given by threshold (α_i).

of subjects with higher α_i decreases as the quality drops. For example, the fourth bin in Fig. 11(a) has very few orange subjects (i.e., subjects with a high threshold).

FYN does not display the same quality to all subjects for two reasons: (1) the server streams different quality levels for each movie, and (2) subjects impose their own lower limit on the displayed quality. For example, if a subject determines quality below VMAF 80 to be unacceptable, they will press the button whenever VMAF reaches 80 and they will never see lower quality. As the result, the overall quality presented in the FYN experiment depends on subjects. On the other hand, in the CVT experiment, everyone views footage of exactly the same quality. Regardless of personal preference, subjects cannot influence the next sequence.

Fig. 12 shows the distribution of ACR ratings for the CVT experiment. Subjects are color coded according to the CVT sensitivity to quality, Δ_i. We see that the Δ_i influences the distribution (e.g., less orange for bad quality than for excellent), but this influence is more subtle than the color distribution trends in Fig. 11a.

Taking into account that the quality range in CVT is twice that of the range in FYN, and not taking into account the overall distribution, this small influence upon the same subjects indicates they behave differently in CVT and FYN. Four subjects never allowed the FYN quality to drop below

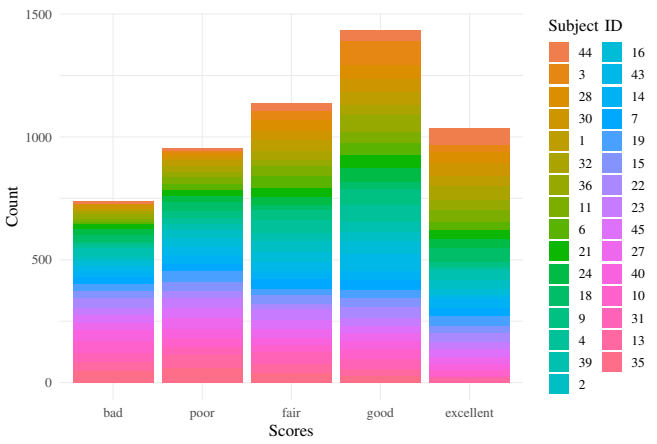


Fig. 12: Distribution of number of answers per rating category in CVT. The subjects are sorted by subject bias (Δ_i).

VMAF 80. We would expect these four subjects to have given an ACR rating of fair or worse to anything below VMAF 80. However, those subjects indicated a quality of good or excellent for 14% to 24% of the answers of CVT sequences with VMAF $<$ 80. This suggests that their actions in FYN differed from their ratings in CVT.

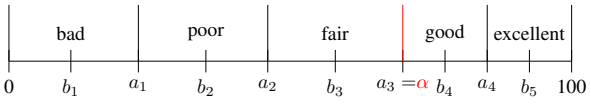
C. Using Sensitivity Threshold to Simulate Rating Distributions

FYN behaviors suggest that subjects in a conventional test do not want to see videos when VMAF drops below some threshold. This threshold probably influences video streaming service subscriptions, the quality of free videos watched at home, and how subjects interpret the ACR scale.

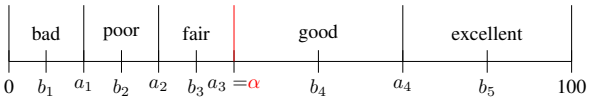
Let us assume that the threshold (α) indicates a fundamental characteristic of each subject. Directly from the FYN experiment, we can estimate α which we call α^{FYN}. If our assumptions about α are true, we should be able to combine our threshold estimate (α^{FYN}) with the expected quality distribution of a subjective test's videos (from VMAF) to model the expected distribution of that subject's ACR ratings. We will need one more piece of information: how α influences the ACR scale. Since we do not know this, we will consider different scenarios.

We will make the following assumptions: 1) subjects want to fix the system's problems any time the quality drops below a threshold α, 2) α differs among subjects, 3) α maps to a specific point on the ACR scale, 4) subject ratings are uniformly distributed across the ACR scale from minimum quality to α, and 5) subject ratings are uniformly distributed across the ACR scale from α to the maximum quality.

The task is to map the 0 to 100 VMAF scale to five rating levels: excellent, good, fair, poor, and bad. We will begin by assuming that α is on the edge between fair and good. Fig. 13 shows us how the VMAF scale would evenly divide into ACR rating intervals, for two example thresholds: $\alpha = 40$ and $\alpha = 70$. Variable a_u divides the rating categories, b_u is set to the middle of each category, and u is one discrete rating level (i.e., 5 = excellent, 4 = good, 3 = fair, 2 = poor, and 1 = bad).



(a) VMAF to ACR mapping for $\alpha = 70$



(b) VMAF to ACR mapping for $\alpha = 40$

Fig. 13: Mapping between VMAF and ACR scale with assumption that each subject threshold maps to the border between "fair" and "good" on the ACR scale ($a_3 = \alpha$).

Thus, VMAF [0, α] are evenly divided into the bad, poor, and fair; while VMAF (α, 100] are evenly divided between good and excellent. Note: a_u, b_u, and α should also have i subscripts to denote the subject, but we will leave these off to simplify the following description.

The moment we assume that ($a_3 = \alpha$), we can calculate $a_u^{a_3}(\alpha)$ and $b_u^{a_3}(\alpha)$[4]. If we assume that α is on the edge between "good" and "excellent," at ($a_4 = \alpha$), we can adjust our calculations and derive a different set of $a_u^{a_4}(\alpha)$ and $b_u^{a_4}(\alpha)$.

The $a_u^{a_3}$ and $b_u^{a_3}$ values are theoretical. Our estimate of $b_u^{a_3}$, called $\hat{b}_u$, is the average VMAF for all videos that receive a rating of u from subject i. For example, for each CVT subject, we will find all videos that were rated excellent = 5 and calculate the average VMAF for that set of videos. This yields $\hat{b}_5$. We will repeat this for $u \in \{1, 2, 3, 4\}$. If the subject never used a rating level, we assign a default value to $\hat{b}_u$. Now we have five theoretical values $b_u^{a_3}(\alpha)$ and five estimations $\hat{b}_u$. By minimizing the square distance between $b_u^{a_3}(\alpha)$ and $\hat{b}_u$, we can estimate α for CVT experiment (with the assumption $a_3 = \alpha$ and this subject), which we will denote α^{CVT,a_3}:

$$\alpha^{\text{CVT},a_3} : \frac{\sum_{u=1}^{5}(b_u^{a_3}(\alpha) - \hat{b}_u)^2}{d\alpha} = 0 \qquad (2)$$

Equation (2) allows us to derive α^{CVT,a_3} from CVT or any ACR experiment. These per-subject α^{CVT,a_3} estimates can be compared with α^{FYN}, the per-subject thresholds extracted from FYN. This scatter plot is presented in Fig. 14.

The 0.38 correlation between α^{CVT,a_3} and α^{FYN} is significant but close to the arbitrarily selected significance level 0.05. The distribution of data around the fit line is not uniform, which indicates a missing factor in the analysis. Closer inspection of this scatter plot shows that there are two different groups. Fig. 15 fits a separate line to each of these two groups. The relaxed group is shifted up and to the left (colored blue) and has lower FYN thresholds on average. The demanding group is shifted down and to the right (colored pink) and has higher FYN thresholds on average.

The division between the Fig. 15 groups is done by hand, but the level to which the results improved is much beyond

[4]For interested readers: for $a_3 = \alpha$ we have: $a_1^{a_3}(\alpha) = \frac{1}{3}\alpha$, $a_2^{a_3}(\alpha) = \frac{2}{3}\alpha$, $a_3^{a_3}(\alpha) = \alpha$, $a_4^{a_3}(\alpha) = \frac{1}{2}\alpha + 50$; $b_1^{a_3}(\alpha) = \frac{1}{6}\alpha$, $b_2^{a_3}(\alpha) = \frac{1}{2}\alpha$, $b_3^{a_3}(\alpha) = \frac{5}{6}\alpha$, $b_4^{a_3}(\alpha) = \frac{3}{4}\alpha + 25$, $b_5^{a_3}(\alpha) = \frac{3}{8}\alpha + \frac{125}{2}$.

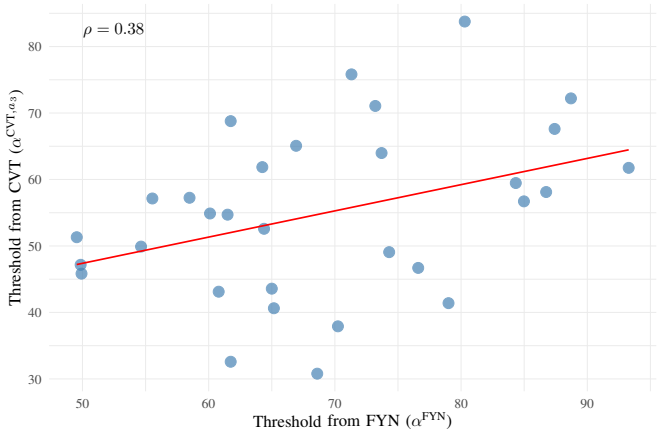


Fig. 14: Scatter plot comparing threshold derived from FYN (α^{FYN}) with threshold derived from CVT (α^{CVT,a_3}) with assumption $a_3 = \alpha$.

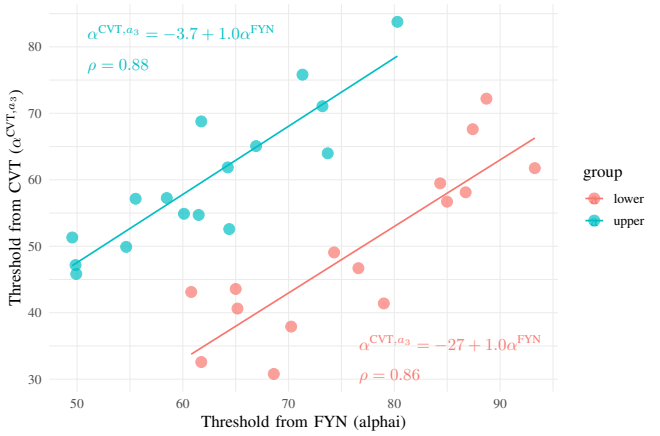


Fig. 15: Scatter plot comparing threshold derived from FYN (α^{FYN}) with threshold derived from CVT (α^{CVT,a_3}). The groups are generated by hand.

what we expected. We have much higher correlations, data are scattered much more uniformly around the fit line, and α^{CVT,a_3} is exactly α^{FYN} just with a shift. For the relaxed group, the shift is small, but for the demanding group the shift is significant.

A simple explanation would be that the demanding group has a different strategy. If these subjects want to watch excellent quality, then it would be reasonable to assume that for them $a_4 = \alpha$. In Fig. 16, the relaxed group assumes that $a_3 = \alpha$ and the demanding group assumes that $a_4 = \alpha$. With this modification, all data are distributed uniformly around a single fit line.

The fit shown Fig. 16 is surprisingly good considering that the FYN measurements of VMAF are imprecise, the VMAF distribution in CVT is not perfectly uniform, and the two experiments are very different. The Pearson correlation is very high ($\rho = 0.89$). We could obtain higher correlation by selecting strategies which give the best fit (up to $\rho = 0.92$), but this is not our goal. Even with our simple strategy, we can see a very strong relation between FYN and CVT.

The existence of different rating strategies was noticed by

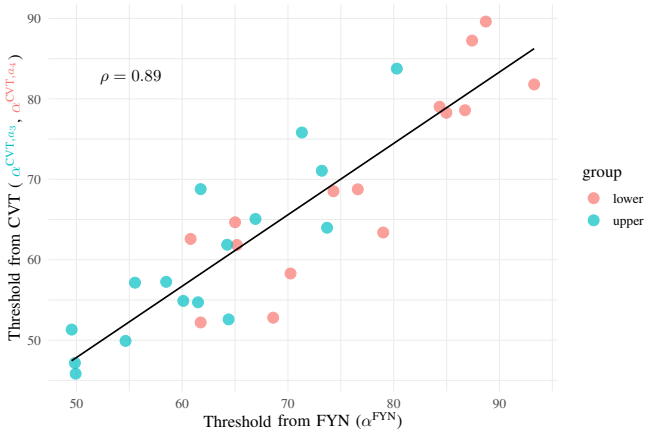


Fig. 16: Scatter plot comparing threshold derived from FYN (α^{FYN}) with threshold derived from CVT, where the two CVT groups use different thresholds: α^{CVT,a_3} for the upper group and α^{CVT,a_4} for the lower group. The groups are generated by hand.

Sackl *et al*. [32]. They also observed premium users (strategy $a_4 = \alpha$, which means "I would like to see excellent quality") and enough quality (strategy $a_3 = \alpha$, which means "I would like to watch good quality") (see Fig. 1). We have no tool to recognize the subject's strategy other than visual inspection of Fig. 14. But knowing this difference in rating behaviors exists, we can extend the conclusions drawn from ACR tests.

VI. Data Analysis: Bitrate Ladder Testing

Now that we understand how FYN compares to a conventional subjective test, we will examine FYN in isolation. This section describes how the FYN experiment design can be used to test adaptive bitrate streaming (ABS) services.

ABS systems divide each content into short segments, each encoded at multiple bitrates. The ABS server chooses which version to send, depending on the available bandwidth and other factors [46]. The most important goal is to prevent stalling, so if the available bandwidth is x, the systems should select a significantly lower bitrate b. Other goals include providing high quality, avoiding rapid changes in quality, creating encodings that span a specific range of bitrates, limiting the number of encodings per segment, and optimizing the other encoding parameters (e.g., rescaling). Limitations on the accuracy and stability of the bandwidth estimations must be taken into account. As always with multiple optimization criteria, it is difficult to find a unique solution. The FYN experiment design could help service providers choose optimal encoding parameters.

Because some FYN subjects did not have a movie preference, we have data on "The Highwayman" from six subjects and data on "Faraway" from nine subjects. This data allows us to compare subject reactions and button pressing probabilities for two ABS ladders, across multiple subjects. Fig. 17 shows the range of responses across these pools of subjects with bootstrap confidence intervals. The line is the average response, and the shaded area is the range of responses. The

red group contains the various other subjects and movies. The fairly smooth response indicates that the server chooses unique VMAF levels for each movie.

Data for “The Highwayman” and “Faraway” distinctly show the impact of three VMAF encoding levels. For “Faraway,” 28%, 29%, and 20% of button presses are associated with only three bitrate levels, at VMAF 58, 67, and 78, respectively. For “The Highwayman,” 28%, 22%, and 20% of button presses are associated with VMAF 52, 67, and 88, respectively. These large clumps of button pressing responses indicate that the quality changes may have been too rapid.

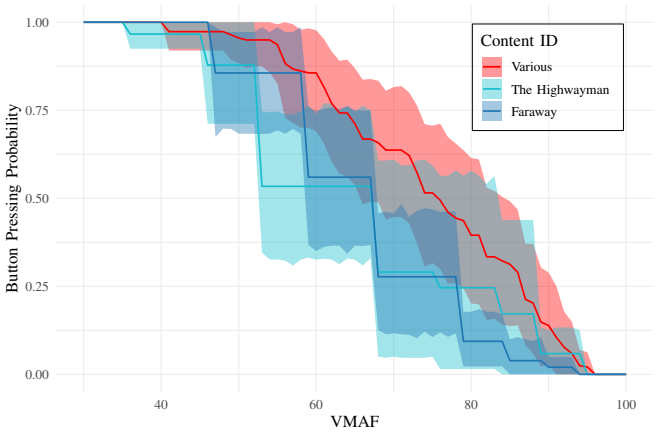


Fig. 17: Bootstrap confidence intervals for button-pressing probability by content ID.

An interesting solution would be limit the press probability to certain level, like 10%, for each encoding level. FYN could be combined with a classic ACR experiment to make the probability prediction more robust and cheaper than running a full FYN experiment. This would be interesting for future research.

With the results we already have, we can predict the global press probability by estimating psychometric function to all the data (see blue line in Fig. 18). The estimated model allows us to predict the differences in VMAF that should be avoided by searching for which VMAF differences change the probability of button pressing by more than x. We can imagine a strategy where the probability change is small for high quality (since most clients will demand high quality) and the probability change is large for lower quality (since only very few clients will tolerate low quality). Fig. 18 illustrates this distribution of ABS quality levels using the red dots. We assumed the first step in the probability drop is 0.02; each following step is 1.3 times that of the previous step but not larger than 0.1.

VII. Future Experiments

The proposed FYN experiment differs significantly from conventional video quality assessment studies. Unlike standard methods, we had to make several decisions without fully knowing how subjects would behave. Certain modifications in the experiment design could help to better interpret the results.

A key limitation of our study is the uncertainty regarding the behavior of some participants (as indicated in Fig. 16). It

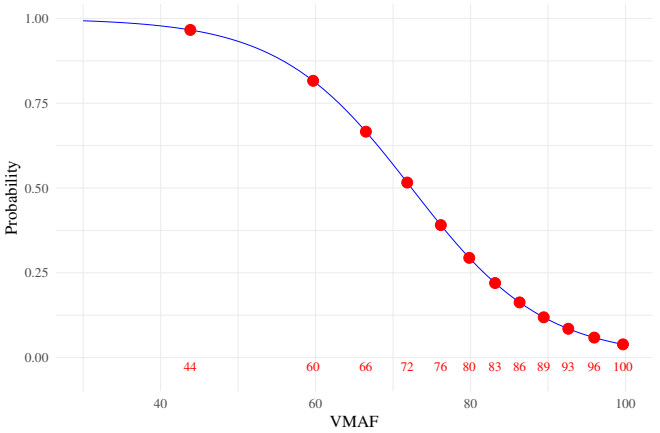


Fig. 18: Button pressing probability for all data as a function of VMAF. The red dots mark a strategy where the quality changes that are most likely to be seen are close together.

is unclear whether subjects from the upper group genuinely required such high video quality or simply pressed the button whenever they noticed any distortion.

To address this ambiguity, a modified setup could include a second button positioned close to the participant. The instructions would specify that subjects should press this button upon noticing any distortion, regardless of its severity, while the primary button would be reserved for cases where the quality is deemed unacceptable. Subjects might be required to press the closer button before pressing the primary button.

Alternatively, the experiment could involve short-form video content, such as YouTube Shorts or TikTok clips. In this scenario, participants would be instructed to press one button when noticing a distortion, or, if the quality is too low, to swipe to the next video to continue watching. Data from a mobile environment would be particularly valuable. Although shorter videos make distortion detection more challenging, a simplified interaction mechanism might offset this, making the overall effect difficult to predict. An important advantage of this approach is that the same video sequences could be reused for a conventional ACR test, enabling direct comparisons.

Another alternative would be to compare ACR and FYN with a third method that has a very different experiment design, such as one using recognition tasks [30], JND [33], completion ratio [47], or the acceptance scale [48]. The differences between these three types of quality assessment experiments could provide insights and help us design a method that more closely mimics people’s behaviors. To the extent possible, the three experiments should use the same subjects and the same video content.

Finally, we suggest an additional modification involving the playback of video sequences with a much finer VMAF granularity—potentially with a version for every single VMAF point. In this configuration, quality degradation would occur more frequently (e.g., every 10–15 seconds) but in smaller steps (1 VMAF point), allowing for more precise estimation of the psychometric function. This enhancement would be particularly useful for designing an accurate bitrate ladder (see

Section VI).

VIII. Conclusions

This paper tackles the important challenge of linking opinion-based subjective experiments to the real world actions of video-service customers. Our initial hypothesis was that this link is weak because of differences between the nature of reporting detectable changes in quality and the nature of acting on quality preferences.

The truth is more nuanced. If the behavioral model we propose is correct, the relationship is strong enough to draw actionable inferences from ACR experiments. If the behavioral model is incorrect, then the direct comparison is weak.

We observed only a weak correlation between a viewer's rating bias (Δ) in the ACR experiment (CVT) and their action threshold (α) in the action-based experiment (FYN). Inter-subject variability was much larger in FYN than CVT. At the extremes, one FYN participant refused to watch any segment below VMAF 91, whereas another viewed an entire film at VMAF 34. One would expect these subjects' ACR scores would differ markedly, yet they did not. Similarly, viewers who never tolerated VMAF < 80 in FYN still labeled 14–24% of CVT clips below that level as "good" or "excellent." Additional studies—such as those outlined in the previous section—are needed to clarify these discrepancies.

Our FYN experiment indicates a large variance between levels of acceptable video quality. By mapping these thresholds to the ACR scale, we uncovered a strong link between FYN and CVT — namely, a subject's acceptability threshold for quality seems to map to the border of "excellent" and "good", or "good" and "fair". Unfortunately, our mapping functions requires data from both ACR and FYN tests, to manually divide subgroups into subgroups with relaxed and demanding threshold behaviors. Future experiments will need to establish an improved mapping function.

Our study also provides practical guidance for bitrate-ladder design in adaptive streaming. This is especially relevant to live streaming, where network impairments are more common and large quality swings must be avoided. FYN data enable more precise ladder tuning than conventional tests. By focusing on specific content or viewer segments, providers can derive distinct curves and build customized ladders.

In summary, FYN introduces a promising experimental paradigm for predicting behavior, not just opinions. Ultimately, behavior, especially within the context of churn, will be a key indicator of the long-term survivability of video streaming platforms.

References

[1] AGH Video QoE Team, "Towards better understanding of factors influencing the QoE by more ecologically-valid evaluation standards," Oct. 2024. [Online]. Available: https://qoe.agh.edu.pl/2020/10/01/towards-better-understanding-of-factors-influencing-the-qoe-by-more-ecologically-valid-evaluation-standards/

[2] ITU-T, *P.910: Subjective video quality assessment methods for multimedia applications*, Geneva, Switzerland, 2023.

[3] ITU-R, *BT.500: Methodologies for the subjective assessment of the quality of television images*, Geneva, Switzerland, 2023.

[4] M. H. Pinson, L. Janowski, R. Pepion, Q. Huynh-Thu, C. Schmidmer, P. Corriveau, A. Younkin, P. Le Callet, M. Barkowsky, and W. Ingram, "The influence of subjects and environment on audiovisual subjective tests: An international study," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 640–651, 2012.

[5] L. Janowski and M. H. Pinson, "The accuracy of subjects in a quality experiment: A theoretical subject model," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2210–2224, 2015.

[6] Z. Li, C. Bampis, L. Janowski, and I. Katsavounidis, "A simple model for subject behavior in subjective experiments," *Proc. IS&T Int'L Symp. on Electronic Imaging: Human Vision and Electronic Imaging*, vol. 2020, pp. 131–1, Jan. 2020.

[7] M. H. Pinson, "The precision and repeatability of media quality comparisons: Measurements and new statistical methods," *IEEE Transactions on Broadcasting*, vol. 69, no. 2, pp. 378–395, 2023.

[8] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," *The Netflix Tech Blog*, vol. 6, no. 2, p. 2, 2016.

[9] L. Janowski, M. H. Pinson, G. Wielgus, K. Koniuch, K. D. Moor, and M. D. Gross, "Watching interesting movies instead of short clips did not change ACR ratings," *Forthcoming*, 2026.

[10] W. Robitza, A. M. Dethof, S. Göring, A. Raake, A. Beyer, and T. Polzehl, "Are you still watching? Streaming video quality and engagement assessment in the crowd," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 2020, pp. 1–6.

[11] S. Ickin, K. Wac, M. Fiedler, L. Janowski, J.-H. Hong, and A. K. Dey, "Factors influencing quality of experience of commonly used mobile applications," *IEEE Communications Magazine*, vol. 50, no. 4, pp. 48–56, 2012.

[12] M. H. Pinson, "Technology gaps in first responder cameras," U.S. Department of Commerce, National Telecommunications and Information Administration, Institute for Telecommunication Sciences, Tech. Rep. NTIA Technical Memorandum TM-17-524, May 2017.

[13] ——, "Why no reference metrics for image and video quality lack accuracy and reproducibility," *IEEE Transactions on Broadcasting*, vol. 69, no. 1, pp. 97–117, 2023.

[14] A. Cohen, "5 KPIs to measure video quality of experience & why they matter," https://www.compiralabs.com/post/5-kpis-to-measure-video-quality-of-experience, Nov. 2019, accessed: 2/28/2025.

[15] D. Nield, "How to make sure you're getting the best streaming quality," https://www.wired.com/story/how-to-get-best-streaming-quality/, Jul. 2022, accessed: 2/28/2025.

[16] B. Frost and L. Jones, "Cable vs. streaming TV: Which is better?" https://www.cabletv.com/blog/cable-vs-streaming, Mar. 2024, accessed: 2/28/2025.

[17] J. Minor and K. Key, "The best live TV streaming services for 2025," https://www.pcmag.com/picks/the-best-live-tv-streaming-services, Jan. 2025, accessed: 2/28/2025.

[18] A. Durrani and S. Allen, "Top streaming statistics," https://www.forbes.com/home-improvement/internet/streaming-stats/, Aug. 2024, accessed: 2/28/2025.

[19] J. K. Willcox, "Guide to streaming video services," https://www.consumerreports.org/electronics-computers/streaming-media/guide-to-streaming-video-services-a4517732799/?msockid=18236b69dfd96f6d19cc7e78de006ef9, Jan. 2025, accessed: 2/28/2025.

[20] D. Rowan, "Nielsen's state of play report reveals that streaming is the future, but consumers are currently overwhelmed by choice," https://www.nielsen.com/news-center/2022/nielsens-state-of-play-report-reveals-that-streaming-is-the-future-but-consumers-are-currently-overwhelmed-by-choice/, Apr. 2022, accessed: 2/28/2025.

[21] "After a boom year in video streaming, what comes next?" https://www.pwc.com/us/en/services/consulting/library/consumer-intelligence-series/consumer-video-streaming-behavior.html, 2021, accessed: 2/28/2025.

[22] R. Thangavel, "Streaming satisfaction report evolving perceptions of value: The shifting sands of SVOD," https://whipmedia.com/wp-content/uploads/2022/06/US-Streaming-Satisfaction-Report-2022.pdf, Jun. 2022, accessed: 2/28/2025.

[23] L. Yang, G. Liu, S. Li, J. Zhao, and T. Jiang, "Environment information enhanced neural adaptive bitrate video streaming for intercity railway," *IEEE Transactions on Broadcasting*, pp. 1–13, 2025.

[24] M. Nguyen, D. Lorenzi, F. Tashtarian, H. Hellwagner, and C. Timmerer, "DoFP+: An HTTP/3-based adaptive bitrate approach using retransmission techniques," *IEEE Access*, vol. 10, pp. 109 565–109 579, 2022.

[25] Z. Wan, B. Zhang, D. Chen, P. Zhang, D. Chen, J. Liao, and F. Wen, "Bringing old photos back to life," in *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, 2020, pp. 2747–2757.

[26] Z. Wan, B. Zhang, D. Chen, and J. Liao, "Bringing old films back to life," *CVPR*, 2022.

[27] C. PT and C. Stacy. Top 100 most viewed YouTube shorts of all. Accessed: 1/12/2026. [Online]. Available: https://www.youtube.com/playlist?list=PLNc3mV34rpBkX50Sng9WdWFxIFEFua-f4

[28] J. Aldredge. (2024, Nov.) Instagram lowers the quality of less popular content. Here's what creators can do about it. Accessed: 1/12/2026. [Online]. Available: https://www.owc.com/blog/instagram-is-reportedly-lowering-the-quality-of-less-popular-content-heres-what-creators-can-do-about-it

[29] H. Yan, H. Fu, Y. Li, T.-H. Lin, G. Wang, H. Zheng, D. Jin, and B. Y. Zhao, "On migratory behavior in video consumption," *IEEE Transactions on Network and Service Management*, vol. 18, no. 2, pp. 1775–1788, 2021.

[30] M. I. Leszczuk, I. Stange, and C. Ford, "Determining image quality requirements for recognition tasks in generalized public safety video applications: Definitions, testing, standardization, and current trends," in *2011 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, 2011, pp. 1–5.

[31] ITU-T, *P.912: Subjective video quality assessment methods for recognition tasks*, Geneva, Switzerland, 2008.

[32] 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 *2012 Fourth International Workshop on Quality of Multimedia Experience*, 2012, pp. 170–175.

[33] J. Zhu, A.-F. Perrin, and P. L. Callet, "Subjective test methodology optimization and prediction framework for just noticeable difference and satisfied user ratio for compressed HD video," in *2022 Picture Coding Symposium (PCS)*, 2022, pp. 313–317.

[34] S. Nami, F. Pakdaman, M. R. Hashemi, and S. Shirmohammadi, "BL-JUNIPER: A CNN-assisted framework for perceptual video coding leveraging block-level JND," *IEEE Transactions on Multimedia*, vol. 25, pp. 5077–5092, 2023.

[35] J. Zhu, P. Le Callet, A.-F. Perrin, S. Sethuraman, and K. Rahul, "On the benefit of parameter-driven approaches for the modeling and the prediction of satisfied user ratio for compressed video," in *2022 IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 4213–4217.

[36] J. Liu, J. Zhu, and P. Le Callet, "Bridge the gap between visual difference prediction model and just noticeable difference subjective datasets," in *2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2023, pp. 1–5.

[37] M. Zhang, B. Gao, G. K. Groth, D. Hermann, and K. Brunnström, "Digital rear view mirrors with augmented reality in comparison with traditional rear-view mirrors," in *IS and T International Symposium on Electronic Imaging Science and Technology*, vol. 36, no. 11. The Society for Imaging Science and Technology, 2024.

[38] K. Brunnström, M. Sjöström, M. Imran, M. Pettersson, and M. Johanson, "Quality of experience for a virtual reality simulator," in *HVEI*, 2018, pp. 1–9.

[39] Netflix. (2022) Netflix open content. Accessed: 1/12/2026. [Online]. Available: https://opencontent.netflix.com/

[40] S. Göring, R. R. Ramachandra Rao, S. Fremerey, and A. Raake, "AVrate Voyager: an open source online testing platform," in *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2021, pp. 1–6.

[41] AVRateNG. AVRateNG – GitHub project. Accessed: 1/12/2026. [Online]. Available: https://github.com/Telecommunication-Telemedia-Assessment/avrateNG

[42] K. S. Durbha, H. Tmar, C. Stejerean, I. Katsavounidis, and A. C. Bovik, "Bitrate ladder construction using visual information fidelity," in *2024 Picture Coding Symposium (PCS)*, 2024, pp. 1–4.

[43] V. Heldestad Lilliesköld and E. Nordh, "Method-of-limits; Cold and warm perception thresholds at proximal and distal body regions," *Clinical Neurophysiology Practice*, vol. 3, pp. 134–140, 2018.

[44] D. Linares and J. López-Moliner, "quickpsy: An R package to fit psychometric functions for multiple groups," *The R Journal*, vol. 8, no. 1, pp. 122–131, 2016. [Online]. Available: https://doi.org/10.32614/RJ-2016-008

[45] F. A. Wichmann and N. J. Hill, "The psychometric function: I. Fitting, sampling, and goodness of fit," *Perception & Psychophysics*, vol. 63, no. 8, pp. 1293–1313, 2001.

[46] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A survey on quality of experience of HTTP adaptive streaming," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 469–492, 2015.

[47] P. Lebreton and K. Yamagishi, "Study on viewing completion ratio of video streaming," in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, 2020, pp. 1–6.

[48] A. Ak, P. Le Callet, A. Gera, H. Tmar, D. Noyes, and I. Katsavounidis, "Sustainable video streaming using acceptability and annoyance paradigm," in *2024 32nd European Signal Processing Conference (EUSIPCO)*, 2024, pp. 952–956.