A SYSTEMATIC APPROACH TO VIDEO QUALITY ASSESSMENT AND BITRATE PLANNING

Sean T. McCarthy, Ph.D.
ARRIS

Abstract

We present a streamlined method of setting operational video quality and bandwidth using either subjective or objective testing, using individual golden-eyes or focus groups of any size. A key feature of the method we present is it provides a means of validating test content to ensure that it is representative of actual programing. Another key feature is that this method enables bandwidth planning based on both average video quality and the probability that hard-to-compress content will fall below a predetermined just-acceptable video quality threshold. The data and analysis we present are intended to aid in planning video quality and bandwidth resources across a range of service offerings from OTT through Ultra HD.

INTRODUCTION

Cable operators and other MVPDs can significantly reduce the complexities of providing ever more viewing options to subscribers by streamlining their approaches to setting video quality thresholds and the minimum bitrates for achieving those thresholds across all streams.

As pay TV providers continue to pursue ways to improve bandwidth efficiency through better encoding techniques on their legacy SD and HD channels, they must also be able to set new quality and bandwidth benchmarks in response to subscriber demand for access to premium content on every type of IP-connected device. This entails working with multiple adaptive bitrate (ABR) streaming modes to determine minimum bitrates for delivering content over fixed and wireless outlets at acceptable quality levels for each type of device.

At the same time, service providers are exploring how to benefit from the introduction of HEVC (High Efficiency Video Coding) compression technology in the IP and legacy domains. This includes preparations for migration to the next plateau in display resolution, 4k Ultra HD. Quality assessment processes are integral to these efforts as well.

These developments present unprecedented challenges to performing systematic, scientifically valid quality assessment at a time when content licensing policies and new approaches to monetization require consistent cross-platform quality performance. Cable and other network operators must be able to define acceptable viewing parameters for every type of service and set the minimum bitrate requirements for whichever encoding protocols they’re using to deliver those services.

As part of this process they must be able to comparatively determine which encoding solutions achieve the best results. And their procedures must include a straight-forward approach to determining whether the technical nuances of the test files they use in focus groups and other testing forums mirror the full scope of nuances in motion, color, zooms, and other dynamics to be found in real programming.

Meeting these challenges requires a practical approach to video quality assessment that is fast, rigorous and cost-effective. This can’t be done simply by relying on objective methods like PSNR (Peak Signal-to-Noise Ratio), which were not designed to gauge the actual human viewing experience[1]. Yet there is nothing quick or cost-effective about repeatedly going through all the time-consuming and costly steps associated with a comprehensive approach to the subjective MOS (Mean Opinion Score) method of assessing real viewing experience, as prescribed in guidelines set by the ITU's BT.500 video assessment recommendations[2].

Instead, operators must be able to take a much more streamlined approach to perceptual video quality assessment which still can be relied on to generate accurate results from the MOS method of measuring real human experience while affording operators the flexibility to employ objective measures wherever appropriate to specific assessment goals. This simplified approach should generate quantifiable results that can be easily communicated in data-driven discussions about video quality on all services. This means the methodology must be sufficiently comprehensive to generate a matrix of average test scores at multiple bitrates across a wide range of video sequences representing the gamut of variables in each service category.

By compiling results from all test sequences in a systematic, quantifiable way operators will be able to use analytic techniques to quickly find answers to different types of questions. For example, operators should be able to use the scores from average viewing experiences at various bitrates across multiple video test sequences to calculate optimal quality thresholds. Looking at test scores generated at different levels of complexity in the video test sequences, they should be able to calculate what the operational bitrate threshold should be for achieving the target quality for a given class of video service.

Operators also want to be able to identify the sequences that have the most negative impact on overall quality performance so that they can focus on finding solutions that will do the most good in raising quality levels for any given bitrate or, alternatively, lowering the bitrate for achieving a given level of quality. They want to be able to identify and quantify the differences in results generated by different encoding systems across multiple test sequences at multiple bitrates. They want to determine whether improvements enabled by new solutions are of sufficient magnitude to merit replacing solutions currently in use.

As explained in the ensuing discussion, extensive commercial experience with a wide range of quality assessment procedures has led to development of a scientific approach to video quality assessment and bitrate planning that will allow operators to achieve highly reliable results in far less time and at far less expense than was previously the case. This is not a one-size-fits-all approach. It provides leeway for using various combinations of subjective and objective testing with reliance on scoring from one or more outside focus groups or from just one individual "golden eye" on the operator's staff.

In all cases, the method involves a means of validating test content to ensure it not only represents the full scope of video characteristics in actual programming but also introduces uncommon extremes that serve to stress test encoders. The method also supports bandwidth planning based on average video quality while accommodating the probability that hard-to-compress content will occasionally fall below a predetermined, just-acceptable video quality threshold.

NEW CHALLENGES IN VIDEO QUALITY ASSESSMENT

From a cable perspective, not very long ago the tasks associated with defining minimum levels of acceptable video quality and the bitrates required to hit those targets focused on just two types of video streams – MPEG-2-encoded SD and HD programming delivered over QAM channels to set-top boxes. Today, operators must also be able to perform quality assessment and bitrate planning on those legacy channel streams when MPEG-4 H.264 encoding is used. And they must be prepared to apply similar assessment procedures to HEVC H.265 encoding as well.

HEVC and 4k UHD

Encoding methodologies applied with MPEG-2 and MPEG-4 are relatively stable, which means operators need to occasionally look at improved encoding techniques to determine whether such improvements translate into gains in bitrate reduction that justify the costs of replacing existing systems. But, where HEVC is concerned, the protocol is still in a relatively unstable phase of implementation, which means assessments of video quality for purposes of setting bitrates will have to be made more frequently in the early rollout phases and beyond.

This is especially true in the use of HEVC to deliver 4k UHD services, which will most likely be a primary driver to early use of HEVC in cable and other MVPD networks. The emergence of 4k UHD requires the setting of new quality thresholds and bitrates. Given the increase in bitrates required for 4k even with the use of HEVC, incremental improvements in encoders will be even more significant to operators than has been the case with SD and HD encoding. As operators add ever more content to their 4k service portfolios, updated video quality assessment will be vital to ensuring maximum bandwidth efficiency as HEVC matures.

Streamed IP Video

Greatly complicating matters is the need to deliver premium video over ABR streams to smartphones, tablets, computers, game consoles, IP set-tops, and smart TVs over Wi-Fi and mobile 3G and 4G connections. This introduces new form factors requiring different bitrate settings to reach minimum quality thresholds, depending on which encoding protocols are in use.

To determine what the appropriate minimum quality thresholds should be for each category of device requires that operators be able to quantify measured video quality in a way that accurately correlates with the quality of actual human viewing experience in accord with widely accepted parameters. In so doing they can determine what the minimum bandwidth requirements will be for a given range of bandwidth contention levels across multiple types of devices, thereby avoiding excessive spending on network capacity as they seek to accommodate the multiscreen IP service paradigm.

Complicating matters is the fact that configurations required for ABR distribution are different from standard broadcast configurations. Interlaced broadcast video must be converted to progressive scan. Frame rates and horizontal and vertical resolutions must be adjusted to comport with the types of devices targeted by the ABR streams. How these processes are performed in today’s automated transcoding systems can impact the quality of the viewing experience and so must be accounted for in analyzing the results from tests used in video quality assessment.

OVERVIEW OF VIDEO QUALITY ASSESSMENT METHODS

Much attention has been devoted to video quality assessment over the decades. There are 2 main kinds of tests: subjective and objective. Subjective methods use systematic and repeatable psycho-visual tests to quantify human judgment of video quality. Objective methods rely on computer-based algorithms to attempt to predict human judgment.

When testing video quality, it is important to understand that “video quality” is an imprecise term. What is really important is being able to answer specific actionable questions such as:

- "What is the minimum operational bitrate that will provide a good video quality of experience to my subscribers?"
- "How often might my subscribers notice problematic video quality?"
- "Is encoder A better that encoder B?"
- "Did the latest encoder upgrade improve video quality or allow a lower operational bitrate?"
- "What is standing in the way of better video quality and lower bitrates?"

For all video quality assessment methods – subjective and objective – the reliability of the results obtained depends crucially on the extent to which the test clips represent the full range of characteristics common to video content. Thus, there need to be clips devoted to emphasizing high motion, random motion, horizontal and vertical pans, zooms and so on. Moreover, the clips chosen from test libraries should go beyond conveying the nuances of these characteristics as they appear in average programming by creating stress conditions suitable to testing the limits of encoder performance.

Subjective Methods

Subjective methods are the golden standard by which all other methods of video quality assessment must be compared. Yet, there is not just one type of subjective test. Which test should be used depends on the question the experimenter hopes to answer.

There are two international standards that give guidance on subjective tests and the methods that should be used to analyze results. ITU-R BT. 500 is a formal recommendation for television. ITU-T P.910[3] is explicitly intended for multimedia applications such as video conferencing, telemedicine, etc.; but may also be applied to television (particularly as the lines between traditional TV and Internet video blur). The aim of all standardized tests is to produce quantitative data such as Mean Opinion Scores (MOS).

Standardized test methods may be classified in various ways. Some display test video in simultaneous pairs either side-by-side or top-to-bottom on one or two displays. Other tests display video test sequences sequentially, thus relying on a viewer's visual memory. An important way in which subjective methods may be classified is based on the manner in which pristine reference video is introduced. In some testing protocols, the reference is "explicit" and the viewer knows which video sequence is the reference. In other protocols, the reference is "hidden," meaning that the viewer does not know which video sequence is the reference. When no reference video sequence is provided, the test may be said to use "implicit" reference, which is usually based on the test viewers' previous experience and expectations with respect to the video quality of commercial television.

The standardized method called Absolute Category Rating (ACR) is worth highlighting as it inspires the method we describe in this paper. The ACR method is a sequential, implicit-reference subjective method. Short test video sequences are displayed and the viewer(s) scores the quality of the video after each test sequence. Each individual test sequence is a particular short clip (10 seconds is a good length) processed or compressed to a different extent. The advantage of ARC is that it is relatively easy to set up. It is also similar to the way in which people watch television. People never see and compare television programming to a pristine reference. A potential disadvantage of ARC is that special care needs to be taken to create a range of test video sequences that include discontinuities such as scene changes, fades, graphic scrawls, etc. typically seen in commercial television.

In addition to the subjective testing methods specified by international standards, there are non-standardized methods that are routinely used to make business decisions involving multi-million dollar equipment purchases. One example of a non-standard method is the “side-by-side bake-off” in which two versions of test video are compared on side-by-side displays. Often the evaluator will go “eye to the glass” to inspect tiny artifacts at very close distances, much closer than any usual viewer. The advantage of this approach is that it is quicker and easier to implement than any standardized method. It also enables an evaluator to compare different video processing options – often products from competing vendors – to highlight and isolated differences with great precision. The disadvantages are that it does not produce quantitative data nor does it give a good indication of how a subscriber or other viewer might react to the artifacts.

It is important to recognize that subjective testing creates an unnatural television viewing experience that can impact how the results should be interpreted. When someone is watching television for the sake of watching television, that person is not actively looking for artifacts. Rather, the viewer notices poor quality when the artifacts are significant enough to grab the viewer’s attention. In subjective testing, a viewer’s tasks are to inspect the television display(s), mentally classify the importance of the any distortions that can be seen, and physically record scores. Thus, subjective testing provides information that marks the boundary of what a viewer could see rather than providing direct information about what a viewer would see during normal television viewing.

Objective Methods

Subjective testing can never be as fast and automatable as objective computer-based algorithms. That is the major advantage of objective testing and why it is so attractive and tempting to use instead of subjective testing.

Objective methods and metrics may be classified as either fidelity-centric or perceptual-centric. An objective metric can be said to be useful if it is accurate and precise, is consistent across different kinds of content, and varies monotonically with subjective scores.

Fidelity-centric methods are the quickest and easiest to use. They quantify the numeric difference between processed and unprocessed video sequences. Peak signal-to-noise ratio (PSNR), sum of absolute differences (SAD), and mean-squared error (MSE) are very common fidelity-centric metrics that accumulate pixel-by-pixel differences. The Structural SIMilarity (SSIM) index[4] and the related multiscale-SSIM (MS-SSIM) analyze correlations between groups of pixels in a test and reference. SSIM and MS-SSIM are moderately more computationally demanding than PSNR but tend to be better predictors of subjective scores and have thus been gaining in popularity in recent years.

Perceptual-centric methods are those that model human vision. Research and development into perceptual models has been very active for decades and is ongoing. Among the better known and more commonly used perceptual-centric methods [5-9] are: Video Quality Metric (VQM); Sarnoff JND (JND borrows from the psychophysical term “just noticeable difference”); and Visible Differences Predictor (VPD).

PSNR, SSIM, MS-SSIM, VQM, VPD, and Sarnoff JND are full-reference metrics (FR), which means that test videos are compared to unprocessed reference video. As such, full-reference metrics are analogous to subjective tests that use explicit or hidden references. Consequently, full-reference may be thought of directly addressing the same class of video quality questions: i.e., “How noticeable is the difference between processed and unprocessed video?” rather than “What is the video quality my subscriber will normally experience?"

Precise spatial and temporal alignment between test and reference videos is a significant issue when using full-reference methods. Being off by a single frame or by a single pixel vertically or horizontally can have a huge impact that could lead to erroneous conclusions. Many objective video quality products have automated alignment methods, but misalignment errors can be so significant that close attention should always be paid to it. Particular caution must be used when the reference and test videos are at different resolutions, different interlacing, different frame rates, or there are skipped, dropped, or repeated frames in a test video.

Because of the recognized issues using full-reference metrics, significant research and development has gone into developing reduced-reference (RR) and no-reference metrics (NR). Reduced-reference means that some metadata concerning the reference video is made available to the video quality analyzer rather than the reference video itself. No-reference means that the video quality analyzer has no access whatsoever to the unprocessed reference video. As such, no-reference metrics are most closely analogous to someone watching television in a normal way.

As tempting as objective methods are, they should be used judiciously and the results interpreted cautiously. No full-reference, reduced-reference, or no-reference metric can yet take the place of people looking at video. Recent tests on FR, RR, and NR methods for HDTV by the Video Quality Experts Group (VQEG)[10,11] give hope and suggest that we are on the right path to developing objective metrics and methods than can serve as proxy for people; but VQEG states in its conclusions that "None of the evaluated models reached the accuracy of the normative subjective testing."

Often, perceptual-centric methods and products that use them will report results as "MOS" or differential MOS ("DMOS"). While understandable, it is worth noting (at the risk of being too pedantic) that this terminology can be confusing and lead to wrong conclusions, particularly by business colleagues who do not have a deep understanding of video quality assessment and the limitations of certain methods. We find habitual use of the terms "predicted MOS" and "predicted DMOS" when referring to the results of objective tests to be a safer more accurate approach.

Objective methods and metrics are very useful even if they are not yet able to take over from subjective testing entirely. Objective methods are great for monitoring systems and for regression testing. The key is to be cognizant of the interpretation of the objective data. Objective results can highlight differences and changes that might require action from an operator or equipment vendor, but objective results should not yet be used synonymously with quality terms such as "better" or "worse." No doubt the day for artificial subjective video quality intelligence will arrive, but that day is not today.

STREAMLINING THE SUBJECTIVE EVALUATION PROCESS

While formal ITU-R BT.500 and ITU-T P.910 testing can provide the results operators need to make mission-critical decisions, the general perception is that standardized subjective testing is too time-consuming and costly to be practically useful in all the situations where video quality assessment is required in today's operations. In truth, however, long experience in video quality assessment shows that a much less time-consuming, less costly iteration of standardized MOS testing can produce results comparable in reliability to those of the comprehensive approach described in ITU-R BT.500 and ITU-T P.910.

As illustrated in Figure 2, there are 4 main components of the streamlined methodology we have found useful: 1) pick an intuitive scoring metric; 2) create a benchmark viewing environment; 3) create a library of test sequences; 4) decide on one or more decision metrics.

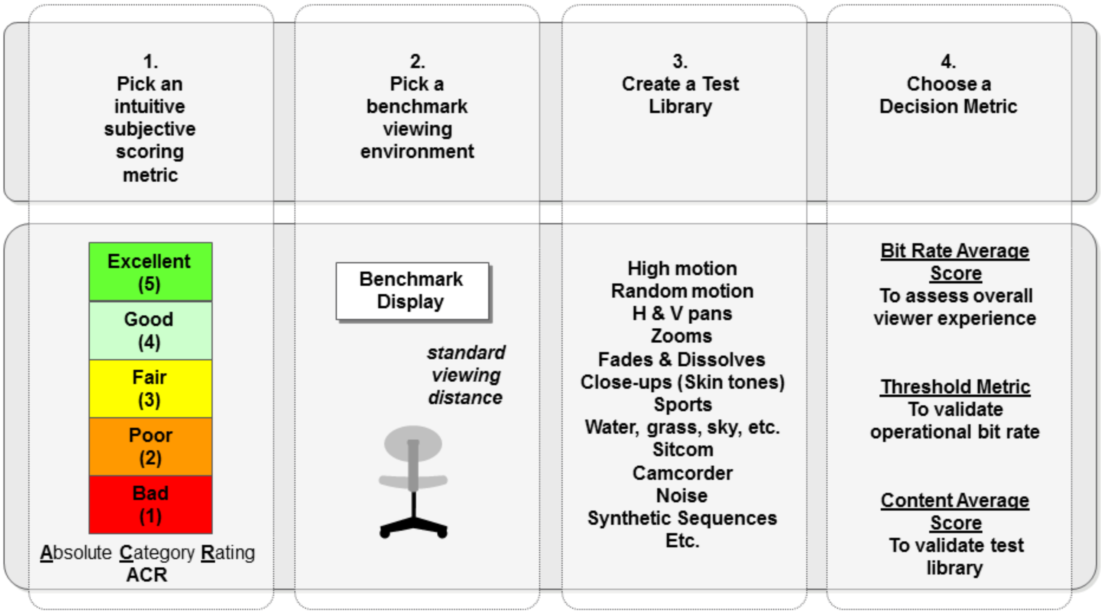


Figure 1. The 4 key components of our streamlined video quality scoring protocol

Simplifying Scoring

Our streamlined method is a simplified form of the Absolute Category Rating (ACR) system specified in ITU-T P.910. We have found that the ARC scoring metric is very easy to explain and most test subjects intuitively understand what is meant by “Bad” (1), “Poor” (2), “Fair” (3), “Good” (4), and “Excellent” (5). The key to choosing a scoring metric is that it not be overly prone to nuance and has stable meaning over time.

Simplifying Viewing Conditions

While the viewing environment must be consistent for all participants in a given testing operation, the conditions that must be factored into creating such an environment are easy to replicate in comparison to the broad set of highly specific conditions such as backlight luminance and minimal ambient noise levels that are described in the ITU recommendations. Test results will be reliable as long as all tests use the same display system with the viewer at a prescribed distance from the screen, room lighting or lack thereof is consistent, and there is no reflected screen glare to interfere with viewing. In our testing, we strive to create viewing conditions that would not be unfamiliar to people watching television at home or viewing content away from home on a mobile device. Although setting up ad hoc testing environments is possible, we have found that having dedicated testing environments, even small ones, helps us be more agile in responding to opportunities to improve our products.

Fewer Test Subjects

The streamlining extends to reducing the number of subjects participating in a test. While there’s always an incremental gain in reliability as the number of subjects increases, those gains become increasingly less significant as the number increases. As long as test viewers are properly vetted to be sure they have good eyesight and no significant perception impairments, good results can be attained from just a handful of viewers. In fact, results obtained from tests involving just a single "golden-eye" viewer on the operations staff whose assessments of video quality have been shown to be consistently reliable can be useful to an operator's decision process.

Building, Validating, and Simplifying a Test Library

Creating a re-usable test library is perhaps the most important aspect of systematic video quality assessment. At a minimum, a test library should represent the variety of content that one would expect to find on commercial television or distributed over one's video service; i.e. "typical" content. However, we have found that also including synthetic and very hard-to-compress "torture" content provides significant advantages.

It is tempting to simply use live programming as input video for video quality testing, particularly for testing transcoders and re-encoding. Using live video has the advantage that, over time, it is representative of one's video distribution service. And it is often as simple to implement as tuning to a live channel. The disadvantage is that it is not repeatable, which is a critical issue for scientifically valid testing. Moreover, live programming rarely provides the "torture" moments for video processing that help define the full range of performance for equipment and services.

We have found that a better approach is to record clips of live programming as either MPEG-2 transport streams or decompressed video for possible inclusion in a master video quality library. Such clips may then be stored and played back in a loop or in concatenation with other test content as many times, and in as many separate tests, as needed.

As mentioned earlier, the test sequences should be chosen to reflect rigorous stress-level representations of all the characteristics that go into assessing video quality. This ensures that the tests will maintain a balance between "easy" and "hard" characteristics, thereby avoiding a skewing of scores too far to the high or low ends of the MOS spectrum. As discussed below, compression-busting "torture" test video can also help to show an operator or equipment vendor what particular technological improvements might be undertaken to improve video quality and/or reduce operational bitrates.

When is a test library complete enough that it can be used to generate actionable data through video quality testing? Validating a test library is very important. It is also important that the test library not become too burdensome. It is often counterproductive to have too many test videos, particularly for subjective testing. It is better to have a smaller set of test videos that strategically spans the range from very simple, through typical, to very hard-to-compress. (Objective methods are a great companion to subjective testing in this regard. They can process larger test libraries to probe for corner cases to be identified for further investigation.)

Extending the Usefulness of Single Stimulus Testing

Most important of all when it comes to simplifying video quality assessment procedures, operators can use the single-stimulus test results to achieve all the goals normally associated with both single- and double-stimulus testing. In other words, they can dispense with the side-by-side and sequential comparative testing between reference and test videos commonly used in encoding vendor "bake-offs" and for determining video quality and bitrate thresholds.

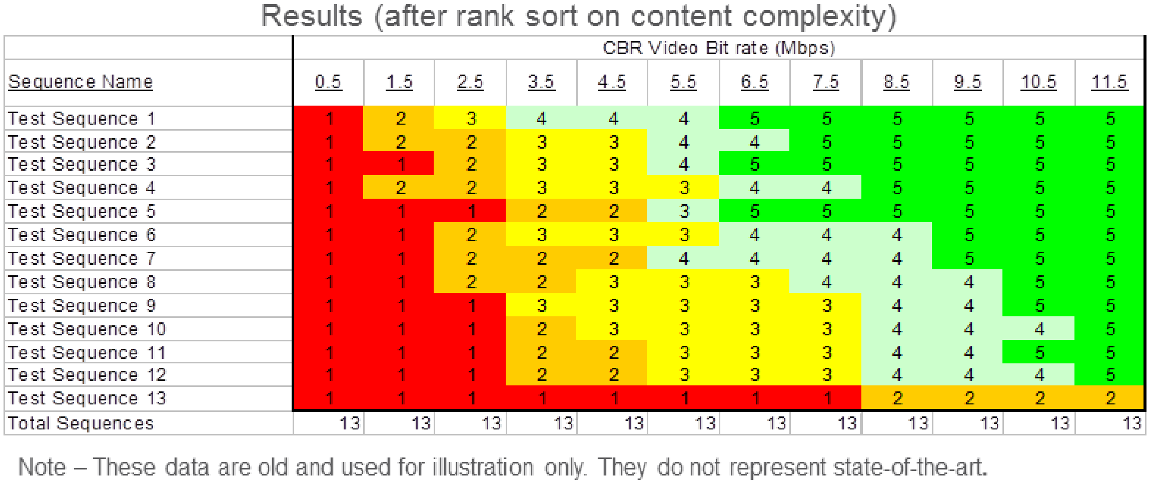
Our Streamlined Testing Protocol

In our streamlined protocol, the subjective testing is carried out by asking a viewer to watch short test clips from test libraries we have developed over time in our interactions with colleagues and customers in cable, telco IPTV, direct-to-home satellite, industry groups, and standards bodies. Test sequences were 10 to 20 seconds each and were displayed up to 3 times in a continuous loop.

After the presentation of each test video, the viewer scored the video quality of the test clip. Each test video represented a different combination of source content and compressed bitrate. Note that our streamlined protocol may be used with test videos that were compressed in either constant bitrate (CBR) mode or variable bitrate (VBR) mode.

Care should be taken in the preparation of test videos. In most of our testing, we play the source video on a short loop and perform continuous real time encoding so as to avoid artifacts related to start-up effects of the rate control algorithm. An alternative approach would be to create a short uncompressed clip from a longer captured compressed bit stream. The uncompressed clip could then be played on a short continuous loop as the test stimulus.

Figure 2 shows data from a subjective testing session from a number of years ago. Although the scores are now obsolete (encoders keep getting better), they provide a useful way to illustrate the data analysis component of our streamline protocol. The data are also somewhat unusual because they are from a "golden eye" session in which a single expert viewer recorded individual scores. This particular subjective testing session was undertaken to provide information with respect to choosing a particular technology development path that required a quick response on our part. We provide this use case here to highlight that we believe our method has value even in such a case, though more viewers will always provide more reliable results.



Results (after rank sort on content complexity)

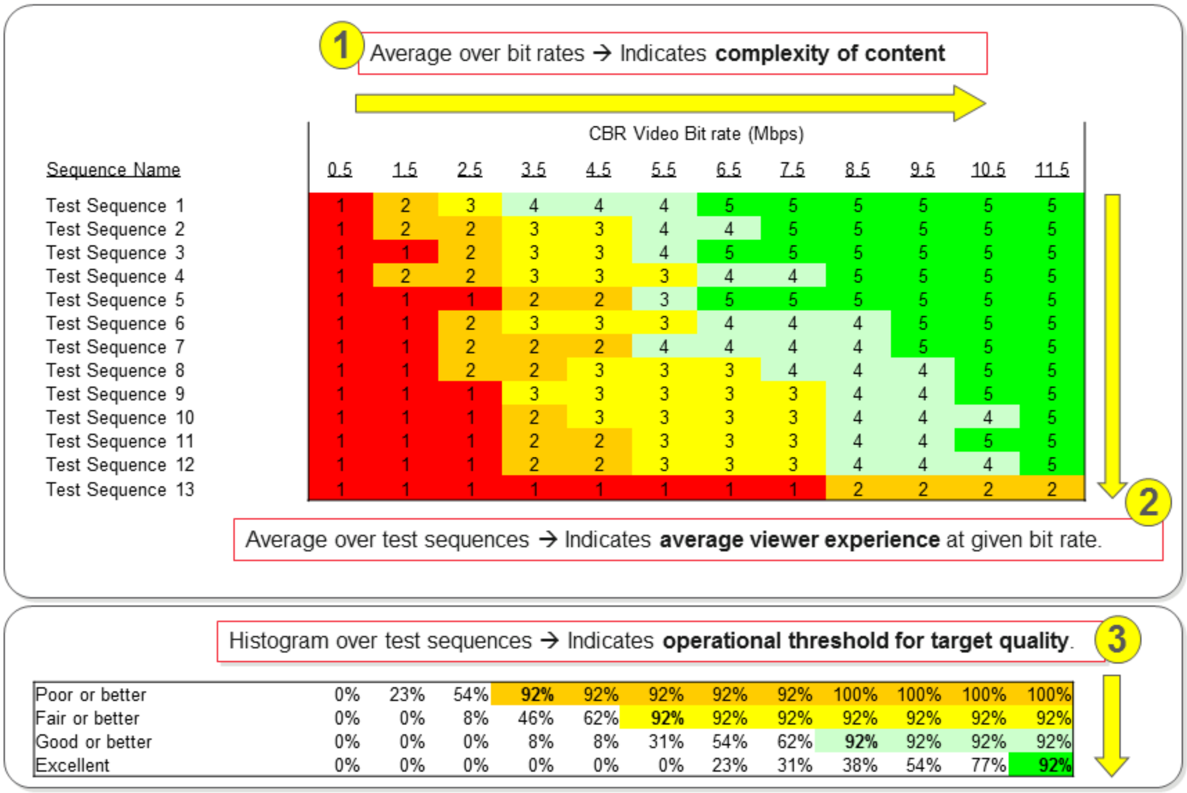
	CBR Video Bit rate (Mbps)											
Sequence Name	0.5	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5	9.5	10.5	11.5
Test Sequence 1	1	2	3	4	4	4	5	5	5	5	5	5
Test Sequence 2	1	2	2	3	3	4	4	5	5	5	5	5
Test Sequence 3	1	1	2	3	3	4	5	5	5	5	5	5
Test Sequence 4	1	2	2	3	3	3	4	4	5	5	5	5
Test Sequence 5	1	1	1	2	2	3	5	5	5	5	5	5
Test Sequence 6	1	1	2	3	3	3	4	4	4	5	5	5
Test Sequence 7	1	1	2	2	2	4	4	4	4	5	5	5
Test Sequence 8	1	1	2	2	3	3	3	4	4	4	5	5
Test Sequence 9	1	1	1	3	3	3	3	3	4	4	5	5
Test Sequence 10	1	1	1	2	3	3	3	3	4	4	4	5
Test Sequence 11	1	1	1	2	2	3	3	3	4	4	5	5
Test Sequence 12	1	1	1	2	2	3	3	3	4	4	4	5
Test Sequence 13	1	1	1	1	1	1	1	1	2	2	2	2
Total Sequences	13	13	13	13	13	13	13	13	13	13	13	13

Note – These data are old and used for illustration only. They do not represent state-of-the-art.

Figure 2. An example of subjective scores for a particular session. In this session, 13 test sequences were each encoded at 12 CBR bitrates from 0.05 Mbps to 11.5 Mbps. Note that the test sequences/bitrate combinations were presented in random order during the test. The data shown here were sorted during data analysis according to average content complexity (see Figure 4.)

Deriving Multiple Answers from a Single Scoring Matrix

As shown in Figure 3, these results can be analyzed in three different ways to provide answers to key questions associated with video quality assessment. One parameter produced in the analytics process is the degree of complexity of each test clip as reflected in the average score at each bitrate. In this example, the least complex clips produce an average score of 5 (“excellent”) at a minimum bitrate of 6.5 Mbps while some of the most complex clips only reach 5 at 11.5 Mbps and one never goes above 2 even at the top bitrate. This analysis suggests that the set of test sequences used in this session represent a range of complexity even in this streamlines testing process.



1 Average over bit rates → Indicates **complexity of content**

Sequence Name	0.5	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5	9.5	10.5	11.5
	CBR Video Bit rate (Mbps)											
Test Sequence 1	1	2	3	4	4	4	5	5	5	5	5	5
Test Sequence 2	1	2	2	3	3	4	4	5	5	5	5	5
Test Sequence 3	1	1	2	3	3	4	5	5	5	5	5	5
Test Sequence 4	1	2	2	3	3	3	4	4	5	5	5	5
Test Sequence 5	1	1	1	2	2	3	5	5	5	5	5	5
Test Sequence 6	1	1	2	3	3	3	4	4	4	5	5	5
Test Sequence 7	1	1	2	2	2	4	4	4	4	5	5	5
Test Sequence 8	1	1	2	2	3	3	3	4	4	4	5	5
Test Sequence 9	1	1	1	3	3	3	3	3	4	4	5	5
Test Sequence 10	1	1	1	2	3	3	3	3	4	4	4	5
Test Sequence 11	1	1	1	2	2	3	3	3	4	4	5	5
Test Sequence 12	1	1	1	2	2	3	3	3	4	4	4	5
Test Sequence 13	1	1	1	1	1	1	1	1	2	2	2	2

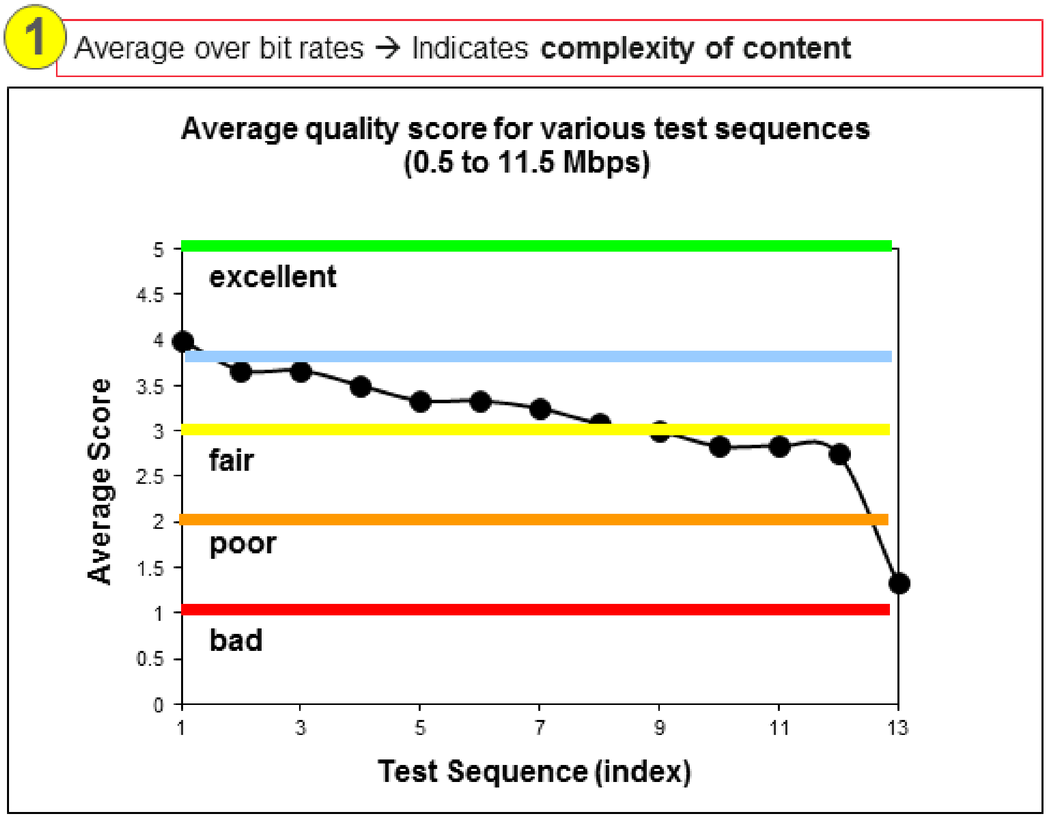
2 Average over test sequences → Indicates **average viewer experience** at given bit rate.

3 Histogram over test sequences → Indicates **operational threshold for target quality**.

Poor or better	0%	23%	54%	**92%**	92%	92%	92%	92%	100%	100%	100%	100%
Fair or better	0%	0%	8%	46%	62%	**92%**	92%	92%	92%	92%	92%	92%
Good or better	0%	0%	0%	8%	8%	31%	54%	62%	**92%**	92%	92%	92%
Excellent	0%	0%	0%	0%	0%	0%	23%	31%	38%	54%	77%	**92%**

Figure 3. An example of analyzing the subjective scores to help answer 3 important technical questions that affect business: 1) “Is the test library sufficiently varied to allow for solid conclusions from the recorded data?” 2) “What may we expect the average viewer experience to be?” 3) “What bandwidth should I plan for in order to achieve a particular level of video quality?”

A more detailed view of the validity of the set of test sequences used may be obtained by calculating the average score across all bitrates for each test clip, as plotted in Figure 4. It is clear that 12 of the 13 test clips cover the just-below-fair to just-above-good range, but only 1 test sequence probes below poor and none above good. The value of plotting the averages in this manner is that it shows us, in hindsight, our test of several years ago would have benefited from a more diverse set of test sequences. If these were results for a test we were performing today, it would be prudent for us to add more test sequences and record more subjective scores. Over time, analyzing the master test library and updating the library based on the results would help to optimize the library so as to support better decisions about video quality and bitrate plans.



1 Average over bit rates → Indicates **complexity of content**

Figure 4. An example of analyzing the subjective scores to determine if the test library is sufficiently varied to allow for solid conclusions from the recorded data

By providing insight into the average viewing experience across all sequences at any given bitrate the methodology allows operators to calculate the overall average score at each bitrate, as reflected in Figure 5. This chart shows that the encoder used in this test produces a good quality of viewing experience, on average, starting at a bitrate of approximately 5.5 Mbps and an excellent video quality starting at around 8.5 Mbps.

These measures can be used by the operator to perform comparative analysis on different encoding systems, as shown in Figure 6. Here the tested performance of the improved encoder shown in Figure 5 is charted with the performance of a previously existing encoder to show key points of disparity between the two, such as the higher bitrates at which the previously existing encoder registers “good” and “excellent” scores.

The third type of analysis provides operators the metrics they need to determine where they want to set the operational bitrate for reaching a targeted level of quality. As shown in Figure 7, the resulting histogram shows the tested encoder achieves fair or better quality on 90 percent of the test sequences at a bitrate of 5.5 Mbps, good or better quality at the 90 percent threshold at a bitrate of 8.5 Mbps and excellent or better at 11.5 Mbps.

Using the same type of histogram results to compare the performance of the two encoders, the operator can assess the degree to which the bitrate threshold for achieving a given quality target can be lowered with use of the improved encoder.

As shown in Figure 8, the threshold can be lowered by a margin of about 1 Mbps for achieving good-or-better performance on 90 percent of the sequences with the improved encoder.

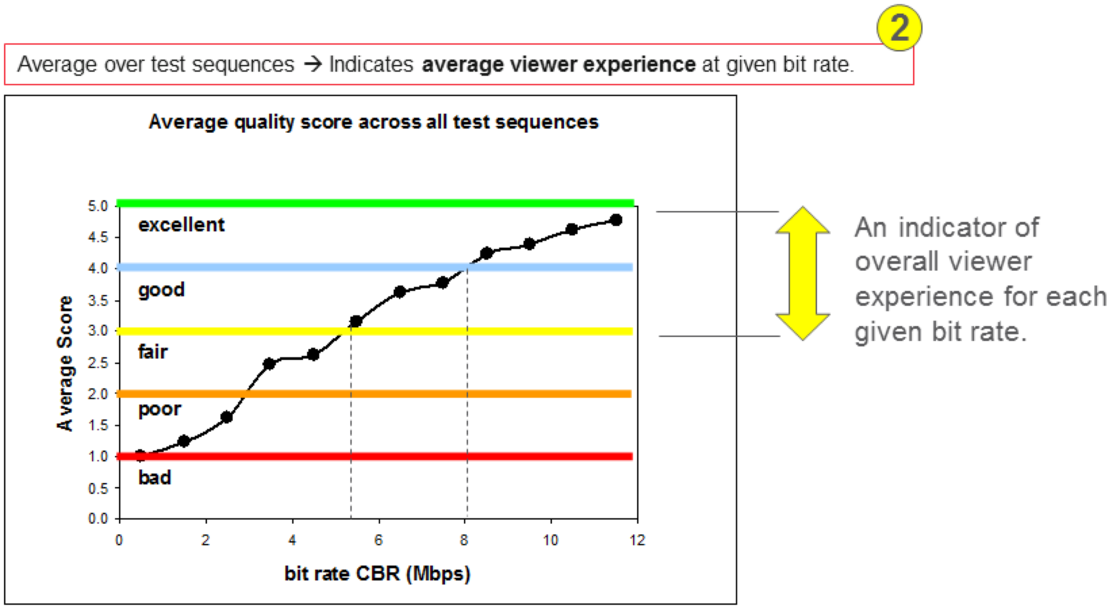


Figure 5. An example of analyzing the subjective scores to provide information that would be predictive of subscriber's overall video quality of experience

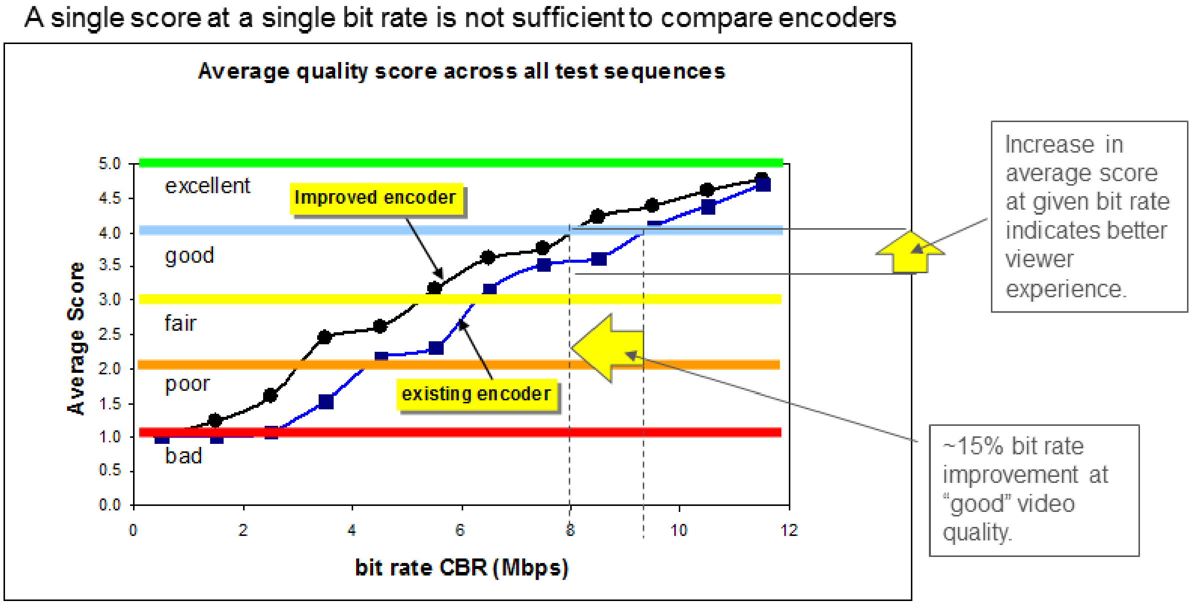
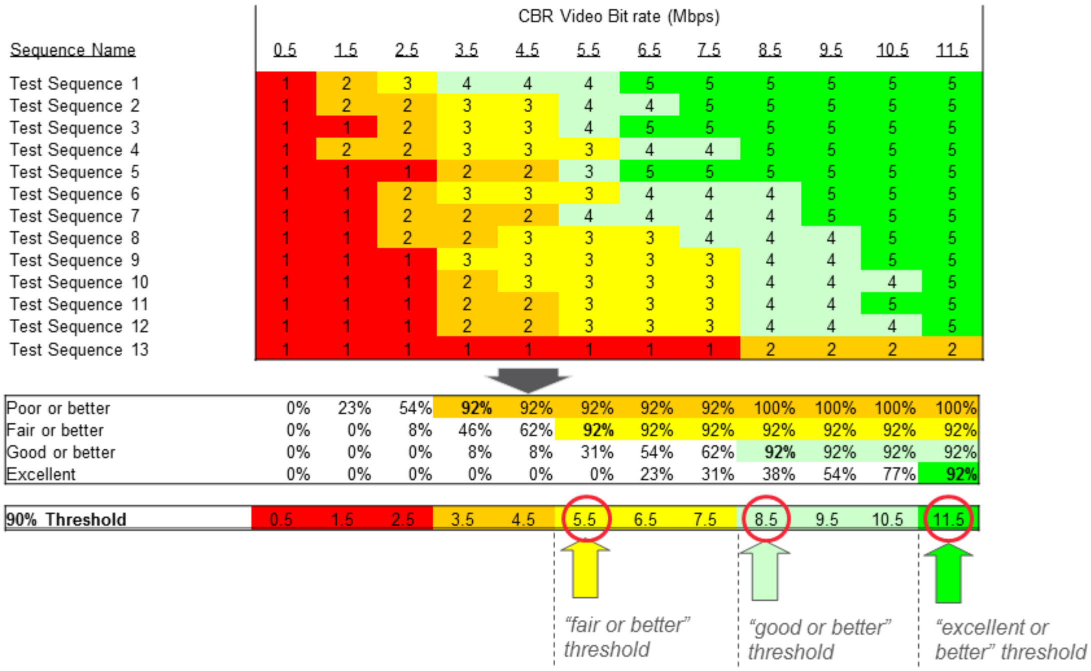


Figure 6. An example quantifying how much one compression option might improve the overall view quality of video experience compared to another compression option



Sequence Name	0.5	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5	9.5	10.5	11.5
Test Sequence 1	1	2	3	4	4	4	5	5	5	5	5	5
Test Sequence 2	1	2	2	3	3	4	4	5	5	5	5	5
Test Sequence 3	1	1	2	3	3	4	5	5	5	5	5	5
Test Sequence 4	1	2	2	3	3	3	4	4	5	5	5	5
Test Sequence 5	1	1	1	2	2	3	5	5	5	5	5	5
Test Sequence 6	1	1	2	3	3	3	4	4	4	5	5	5
Test Sequence 7	1	1	2	2	2	4	4	4	4	5	5	5
Test Sequence 8	1	1	2	2	3	3	3	4	4	4	5	5
Test Sequence 9	1	1	1	3	3	3	3	3	4	4	5	5
Test Sequence 10	1	1	1	2	3	3	3	3	4	4	4	5
Test Sequence 11	1	1	1	2	2	3	3	3	4	4	5	5
Test Sequence 12	1	1	1	2	2	3	3	3	4	4	4	5
Test Sequence 13	1	1	1	1	1	1	1	1	2	2	2	2

Poor or better	0%	23%	54%	**92%**	92%	92%	92%	92%	100%	100%	100%	100%
Fair or better	0%	0%	8%	46%	62%	**92%**	92%	92%	92%	92%	92%	92%
Good or better	0%	0%	0%	8%	8%	31%	54%	62%	**92%**	92%	92%	92%
Excellent	0%	0%	0%	0%	0%	0%	23%	31%	38%	54%	77%	**92%**

90% Threshold	0.5	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5	9.5	10.5	11.5

CBR Video Bit rate (Mbps)
"fair or better" threshold
"good or better" threshold
"excellent or better" threshold

Figure 7. An example of analyzing the subjective scores to help determine how much bandwidth should be planned in order to achieve a particular level of video quality for a certain percentage of programming

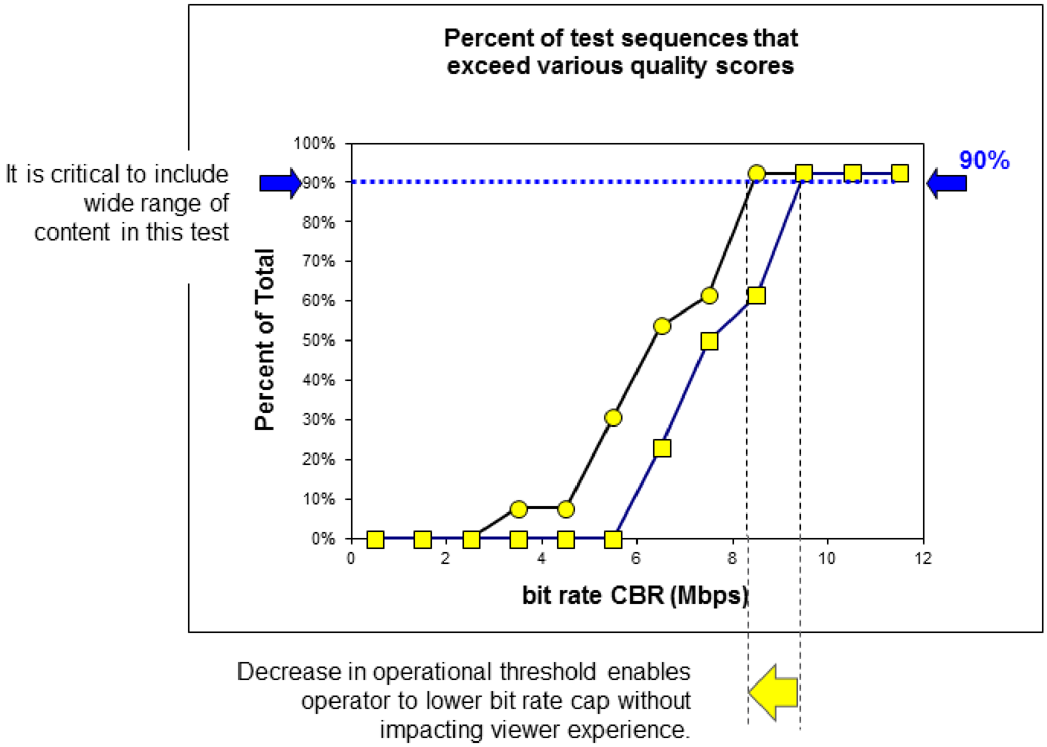
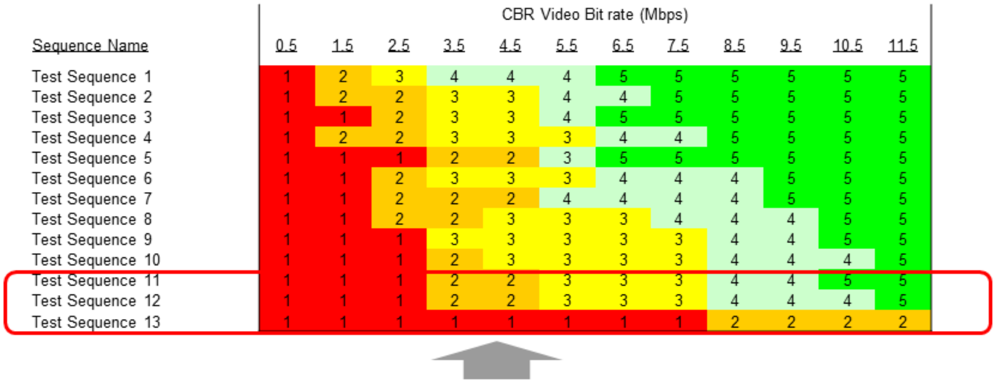


Figure 8. An example quantifying how much one compression option might lower the operational bitrate requirements compared to another compression option

Facilitating Improvements in Performance

The scoring matrix also provides a ready means of identifying the areas of weakest quality performance for a given encoding system, as shown in Figure 9. Whether the matrix is used by encoding designers or operators, they need to be able to quickly see where efforts to improve performance will do the most good.

By focusing on the weakest spots as shown in the test clip scores, developers and operators will be able to obtain the greatest improvement in overall performance ratings. Ongoing efforts to improve performance can be focused on the next most troublesome areas once the performance on the initially targeted set of "troublemakers" is improved as much as technical means allow.



CBR Video Bit rate (Mbps)

Sequence Name	0.5	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5	9.5	10.5	11.5
Test Sequence 1	1	2	3	4	4	4	5	5	5	5	5	5
Test Sequence 2	1	2	2	3	3	4	4	5	5	5	5	5
Test Sequence 3	1	1	2	3	3	4	5	5	5	5	5	5
Test Sequence 4	1	2	2	3	3	3	4	4	5	5	5	5
Test Sequence 5	1	1	1	2	2	3	5	5	5	5	5	5
Test Sequence 6	1	1	2	3	3	3	4	4	4	5	5	5
Test Sequence 7	1	1	2	2	2	4	4	4	4	5	5	5
Test Sequence 8	1	1	2	2	3	3	3	4	4	4	5	5
Test Sequence 9	1	1	1	3	3	3	3	3	4	4	5	5
Test Sequence 10	1	1	1	2	3	3	3	3	4	4	4	5
Test Sequence 11	1	1	1	2	2	3	3	3	4	4	5	5
Test Sequence 12	1	1	1	2	2	3	3	3	4	4	4	5
Test Sequence 13	1	1	1	1	1	1	1	1	2	2	2	2

"The Top 3 Troublemakers"

Focus video quality optimization efforts here.

These 3 test sequences represent the type of content and artifacts that prevent:

- lower operational bit rate targets
- better subscriber quality of video experience

Figure 9. An example of using the subjective scores to help decide where to allocate resources toward improving and optimizing video processing equipment and services

These examples demonstrate how a simplified, systematic approach to single-stimulus MOS testing can enable a data-driven discussion about video quality across all service categories. Such quantifiable results are essential to guiding the selection of quality targets, setting minimum bitrates, choosing encoders, and determining whether to make investments in new encoding solutions.

CONCLUSION

As the need for video quality assessment encompasses an ever-expanding range of cable and other MVPD operations, from legacy SD and HD to TV Everywhere and IPTV to HEVC and 4k UHD, operators cannot afford to base critical decisions on unreliable testing processes. PSNR is not designed to be, nor should it be, used as substitutes for measuring human perception of video quality. Other objective methods leverage knowledge of human vision to a greater or lesser extent, but none of them should yet be relied on as a replacement for people.

Efforts to base key decisions affecting quality thresholds, bitrate settings, and technology choices on real human viewing experience should not be impeded by unnecessary complications in the testing process. There's not enough time or money for executing all the procedures typically associated with using standardized MOS-based measures of human perception as the foundation for decision making.

Fortunately, there is a much simpler, scientifically valid approach to generating and applying the results of MOS testing in video quality assessment. By reducing the number of participants in tests, eliminating overly precise conditions for creating viable test environments and focusing on using the results of single-stimulus testing to answer questions traditionally addressed through more complicated dual-stimulus procedures, operators can perform reliable video quality assessment essential to answering all key questions much faster and at far lower costs.

REFERENCES

1) Z. Wang, A. C. Bovik. "Mean squared error: Love it or leave it? A new look at signal fidelity measures." IEEE Signal Processing Magazine, Vol. 26, No. 1. 2009.
2) ITU. "Recommendation ITU-R BT.500-13. Methodology for the subjective assessment of the quality of television pictures." 2012
3) ITU. "Recommendation ITU-T P.910. Subjective video quality assessment methods for multimedia applications." 2008
4) Z. Wang, A.C. Bovik,, H.R. Sheikh, and E.P. Simoncelli. "Image quality assessment: From error visibility to structural similarity," IEEE Transactions on Image Processing 13: 600–612. 2004
5) Mylène C. Q. Farias (2010). Video Quality Metrics, Digital Video, Floriano De Rango (Ed.), ISBN: 978-953-7619-70-1, InTech, DOI: 10.5772/8038. Available from: http://www.intechopen.com/books/digital-video/video-quality-metrics
6) Wang, Y. (2006). "Survey of objective video quality measurements," Technical Report T1A1.5/96- 110, Worcester Polytechnic Institute.
7) S. Winkler. Digital Video Quality: Vision Models and Metrics. John Wiley and Sons. 2005.
8) S. Winkler "Video Quality Measurement Standards – Current Status and Trends" IEEE ICICS 2009
9) S. Winkler and P. Mohandas. "The Evolution of Video Quality Measurement: From PSNR to Hybrid Metrics" IEEE Trans. Broadcasting Vol 54. No. 3. 2008
10) VQEG. "Final report from the video quality experts group on the validation of objective video quality assessment." March, 2000
11) VQEG. "Report on the Validation of Video Quality Models for High Definition Video Content." June, 2010
12) Winkler, S. "Analysis of Public Image and Video Databases for Quality Assessment? IEEE J. Select Topics Sig. Proc. Vol. 6, No. 6 2012