[Margaret H. Pinson, Lucjan Janowski, and Zdzisław Papir]

VIDEO QUALITY ASSESSMENT

[Subjective testing of entertainment scenes]

TECHNOLOGY PHOTO—© ISTOCK.COM/VIOLETKAIPA
FILM PHOTO—© ISTOCK.COM/F9PHOTOS

This article describes how to perform a video quality subjective test. For companies, these tests can greatly facilitate video product development; for universities, removing perceived barriers to conducting such tests allows expanded research opportunities. This tutorial assumes no prior knowledge and focuses on proven techniques. (Certain commercial equipment, materials, and/or programs are identified in this article to adequately specify the experimental procedure. In no case does such identification imply recommendation or endorsement by the National Telecommunications and Information Administration, nor does it imply that the program or equipment identified is necessarily the best available for this application.)

Video is a booming industry: content is embedded on many Web sites, delivered over the Internet, and streamed to mobile devices. Cisco statistics indicate that video exceeded 50% of total mobile traffic globally or the first time in 2012 and predict that over two-thirds of the world's mobile data traffic will be video by 2018 [1]. Each company must make a strategic decision on the correct balance between delivery cost and user experience. This decision can be made by the engineers designing the service or, for increased accuracy, by consulting users [2].

Digital Object Identifier 10.1109/MSP.2013.2292535
Date of publication: 5 December 2014

Video quality assessment requires a combined approach that includes objective metrics, subjective testing, and live video monitoring. Carefully conducted video quality subjective tests are extremely reliable and repeatable, as is shown in [3, Sec. 8]. This article provides an approachable tutorial on how to conduct a subjective video quality experiment. Our goal is to encourage more companies and universities to perform subjective tests.

A subjective video quality test uses a small set of short video sequences (e.g., 8–20 s) to measure people's opinions of the quality of different video processing options. These tests focus on people's current opinion, as opposed, e.g., to opinions of an entire movie. The goal is to make an impartial judgment about opinion trends. Example applications include choosing between different coding algorithms, comparing one coder at different bit rates, comparing two implementations of the same algorithm, optimizing coder parameters, improving an error concealment algorithm, or selecting a maximum packet loss rate for a service. Video quality subjective tests isolate one factor: video quality. Issues that might confound the experiment data should be excluded [e.g., audio, scene composition, aesthetics, display, environment, device interface, two-way communication, and quality of experience (QoE)].

INTERNATIONAL TELECOMMUNICATIONS UNION RECOMMENDATIONS

The International Telecommunications Union (ITU) recommendations most directly applicable to this tutorial are ITU-R Rec. BT.500 (2012), *Methodology for the Subjective Assessment of the Quality of Television Pictures*; ITU-T Rec. P.910 (2008), *Subjective Video Quality Assessment Methods for Multimedia Applications*; and ITU-R Rec. BT.1788 (2007), *Subjective Assessment of Multiple Video Quality (SAMVIQ)*. ITU-R Rec. BT.500 focuses on video quality and image quality in a home television environment; ITU-T Rec. P.910 focuses on video quality, videotelephony, videoconferencing, and storage/retrieval applications; and ITU-R Rec. BT.1788 identifies one particular rating method. The current version of each recommendation is distributed freely on the ITU Web site (http://www.itu.int/). These procedures remove all distractions from the environment to eliminate variables that might bias the test. The environment is basically an idealized living room: quiet and devoted to this one task. These ITU recommendations assume the reader has some prior knowledge.

The scope of ITU-R Rec. BT.500 is broadcast television and, therefore, entertainment video in either standard-definition or high-definition format. BT.500 specifies highly controlled monitor calibration and lighting conditions (e.g., the ratio of luminance of inactive screen to peak luminance should be ≤ 0.02). The monitor calibration techniques focus on the needs of broadcasters, so an amateur may have difficulty calibrating consumer-grade equipment.

ITU-T Rec. P.910 was designed for video systems at lower bit rates and quality than broadcast television. The wording of P.910's focus may look odd today because terminology has changed. P.910 is appropriate for high-definition television (HDTV) through quarter common intermediate format (QCIF) resolution (176×144). P.910 specifies exact lighting conditions, but they are easier to recreate than BT.500's lighting conditions. The monitor is not calibrated, which is more appropriate for computers, mobile devices, and consumer-grade televisions. ITU-R Rec. BS.1788 is commonly referred to by the acronym of its title: SAMVIQ. This recommendation defines a particular rating scale and method.

Work is underway in the ITU to develop recommendations better suited to new technologies. One example is the newly approved ITU-T Rec. 913 (2014) Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment. This recommendation describes techniques for situations not covered by BT.500 and P.910, including the use of natural lighting or a distracting environment (e.g., a cafeteria or bus). Published at the end of 2014, [51] contains a summary of the differences between the traditional techniques found in this article and P.913.

EXPERIMENT DESIGN

TERMS AND DEFINITIONS

- *Source sequence* (SRC) is the unimpaired video sequence (i.e., the content).
- *Original* refers to the original version of each SRC (e.g., broadcast quality).
- *Processed video sequence* (*PVS*) is the impaired version of a video sequence.
- *Clip* refers to any video sequence, SRC or PVS.
- *Hypothetical reference circuit* (HRC) is a fixed combination of a video encoder operating at a given bit rate, network condition, and video decoder. The abbreviation HRC is preferred when vendor names should not be identified.
- *Full matrix design* consists of n SRCs and m HRCs. All combinations of SRCs and HRCs are included in the experiment for a total of $(n \times m)$ PVSs.
- *Partial matrix design* splits the experiment into two or more smaller matrixes. For example, a two-matrix experiment would have two scene pools (pool A and pool B, with n_A and n_B SRCs, respectively) and two HRC pools (pool A and pool B, with m_A and m_B HRCs, respectively). All combinations of pool A SRC and HRCs are included, plus all combinations of pool B SRC and HRCs, for a total of $(n_A \times m_A + n_B \times m_B)$ PVSs.

GOAL OF EXPERIMENT AND DESIGN CONSEQUENCES

The first goal of a video quality subjective test is to answer a specific question about video encoding, transmission, or decoding. These questions are typically posed as comparisons between one or more variables. The analysis will directly compare pairs of HRCs using identical SRCs, typically via a full matrix design. For example, Younkin and Corriveau [4] use a full matrix design to analyze the impact of playback error severity on quality perception.

The full matrix design allows all HRCs to be directly compared and produces improved accuracy for some analysis techniques (see the section "Choosing a Subjective Scale"). The disadvantage is that less information is obtained about the impact of different source material on the HRCs. This is undesirable because codecs yield very different quality depending upon the scene content, as can be seen in Figure 1, taken from [5]. The x-axis displays the bit rate, and the y-axis displays the mean opinion score (MOS). The boxes and whiskers in Figure 1 show the distribution of PVSs within an HRC. Some of the plotted HRCs span more than half of the absolute category rating (ACR) scale.

Twice as many SRCs can be included in a partial matrix design of two matrices, compared to a full matrix design. This alleviates the subjects' boredom. The partial matrix design allows direct comparisons only of HRCs within an HRC pool. For example, HRCs from pool A cannot be directly compared with HRCs from pool B. Pinson et al. [5] use a partial matrix design to compare the quality of the H.264 and MPEG-2 coders, both with and without packet loss.

Both full matrix and partial matrix designs depend upon two variables: SRC and HRC. A third variable—the environment—is needed to answer questions about interactions between the video signal and the viewing environment. For example, Brunnström et al. [6] explore the relationship between video quality and the viewing angle of the subject to the screen. Thus, Brunnström's HRC definitions specified the viewing angle.

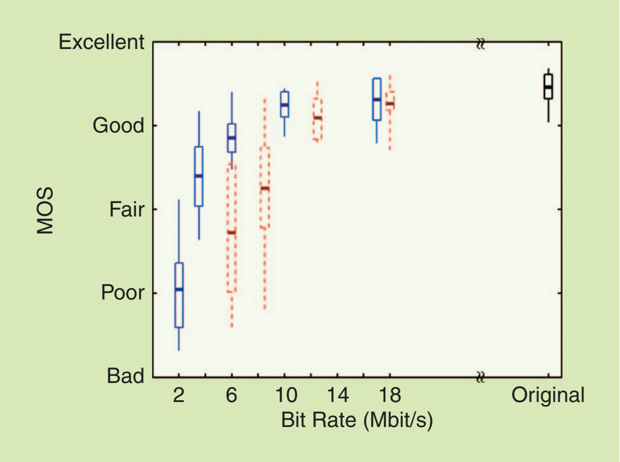
The second video quality subjective test goal is to train a metric or algorithm. For example, an objective video quality metric estimates the quality ratings that would result from a subjective experiment. The accuracy of the resulting metric depends upon the quantity and variety of training data. Thus, the optimal experiment design maximizes the number of SRCs and HRCs for the available number of PVSs. A random pairing of each SRC and a different HRC will accomplish this goal, though most engineers are troubled by the asymmetry. Voran and Wolf [7] provide an example of a subjective experiment designed specifically to train a metric. A full matrix or the partial matrix design is usually less effective, because fewer SRC and HRC can be analyzed. Some experimenters choose a full or partial matrix design anyway, because they want to use the same subjective test for two purposes: to answer a question and to train a metric. Huynh-Thu and Ghanbari [8] provide an example.

The third video quality subjective test goal is to analyze the performance of an existing objective video quality metric or algorithm. The constraint here is not the design of the test—both the full matrix and the partial matrix designs are suitable—but rather the fact that training data cannot be used to test the model's performance. Ideally, this prohibition includes scene content, coder implementations, and coder/network settings (e.g., packet loss rate, and bit rate). Voran and Catellier [9] provide an example of how to design a subjective test to both train and test a metric. This article describes a speech quality experiment; however, the experimental design issues are the same.

Video quality subjective tests can be used in combination with other subjective tests to understand larger quality implications. ITU-T Rec. P.1301 [10] demonstrates this idea for telemeeting systems, and an applied example can be found in [11].

SRC, HRC, AND PVS SELECTION

Video quality subjective tests typically use short sequences (e.g., 8–10 s duration). Pinson et al. [12] provide guidance on choosing a balanced and well-designed set of scenes for a subjective test.



[FIG1] For coding-only impairments, a quality comparison of H.264 (solid blue) and MPEG-2 (dotted red): the box-plot identifies minimum, 25%, mean, 75%, and maximum MOS.

This guidance includes avoiding offensive content, choosing scenes that evenly span a wide range of coding difficulty, deciding whether or not scene cuts should be allowed, and selecting scenes with unusual properties. It is important to use high-quality footage because otherwise the quality impairments in the SRC can obscure any effects of the HRC in the test results. Niu and Liu [13] explain the differences between professional and amateur videography and provide objective criteria for identifying professional video sequences. Amateur footage typically contains aesthetic problems that trigger low video quality ratings (e.g., focus control, color palette, camera motion, shot length, and visual continuity). The Consumer Digital Video Library (www.cdvl.org) provides free downloads of broadcast-quality footage for research and development purposes. Another Web site that offers free footage is http://www.irccyn.ec-nantes.fr/spip.php?article541.

The range of PVS quality should span the scale used to conduct the test (see the section "Choosing a Subjective Scale"). Experiments that contain a narrow range of quality will be frustrating for the subjects and researcher alike since the data are unlikely to show any significant results. It is better to design experiments that span a wide range of quality—or at least a wide enough range that meaningful results can be found.

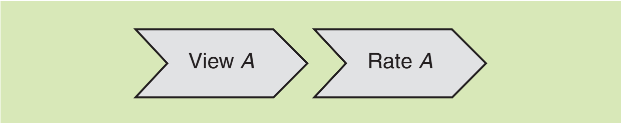
The goal of many subjective experiments is to compare and contrast the quality of various video impairments. This analysis is only possible when HRC creation is limited by two constraints: 1) the definition of each HRC is constant throughout the experiment and 2) if two HRCs are to be compared, then those HRCs must be paired with the same set of SRCs.

The term HRC implies that all PVSs associated with that HRC were created using a constant set of control parameters. A particularly popular HRC definition specifies codec A, profile B, constant bit rate C, and packet loss rate D. As a side effect of this design, difficult-to-encode SRCs will yield a very wide range of quality (excellent to bad), while easy-to-encode SRCs will yield a narrow range of quality (excellent to fair). An alternative approach tries to produce equivalent quality for all PVSs associated with a single HRC. This can be done with variable bit rate encoding or a constant quantization profile value. The problem is that it becomes difficult to reach conclusions about the coder's behavior at different bit rates.

TEST ENVIRONMENT

Traditionally, subjective video quality tests are performed in a controlled laboratory environment. This reduces the effect of extraneous variables on the experiment without requiring a specialized space or great expense. While the potential impact of some elements is debatable, the traditional controlled environment demonstrates your expertise to the research community.

- *Walls*: The walls of the test chamber should be plain white and not show potentially distracting objects (e.g., pictures, clock, and wires). Windows must be covered with light-blocking curtains. Temporary room dividers encourage the illusion of a nondistracting chamber.
- *Floor*: The floor should be a neutral, nondistracting color. Solid gray is traditional.
- *Furniture*: Only necessary furniture should be in the test chamber. The chair provided to subjects should not have wheels. This will encourage the subject to keep a constant viewing distance throughout the test. An upright chair helps to encourage attention on the task.
- *Lighting*: See ITU-R Rec. BT.500 clause 2.1 or ITU-T Rec. P.910 clause 7.1 for lighting conditions. The listed specifications can be met inexpensively using a light meter, full spectrum bulbs, and variable intensity lamp controls.
- *Viewing distance*: See ITU-R Rec. BT.500 clause 2.1 or ITU-T Rec. P.910 clause 7.1 for details. For most experiments, the monitor and chair should be positioned at a defined viewing distance. The viewing distance is traditionally measured in picture heights: four to six times picture heights (H) for standard definition television (i.e., 4H to 6H), 2H to 3H for HDTV, and 8H for smartphones and other very small monitors [14].
- *Monitor*: BT.500 encourages the use of a professional quality monitor to eliminate a potentially confounding variable. For P.910, choose a monitor that matches the application.
- *Background noise*: The test chamber must be quiet, with minimum background noise. If a computer is used to play the videos, the computer should be outside the test chamber.
- *Bystanders*: While a subject is running through the test, the chamber should be used for no other purpose. In some cases, the test chamber will have two or more subjects and the experimenter. People who are interested in seeing the test results come out a certain way should not interact with the subjects, perform the data analysis, or design the test (e.g., product managers).



[FIG2] The ACR rating cycle: the subject watches video clip *A* and then rates *A*.

NUMBER OF SUBJECTS, STIMULI, AND TEST SESSIONS

The reliability of ratings depends upon averaging the data across multiple subjects. While BT.500 recommends a minimum of 15 subjects, a recent study by Pinson et al. [15] endorses a minimum of 24 subjects. Fewer subjects may be used to indicate trending.

Subjects have a limited attention span, and so the typical challenge is fitting all impairments of interest into a set of test sessions that one person can reasonably watch. Preferably, each session should last no more than 20 min, and each subject should spend no more than 1 h rating video. Longer experiments require additional motivation or variety to keep the subjects alert. Payment is the traditional motivator. The best way to add variety into an experiment is to increase the number of SRCs.

Subjective experiments should use at least eight different SRCs. Differences between SRCs are a major variable for every subjective experiment. A large and varied pool of scenes minimizes the risk that the subjective experiment will reach an erroneous conclusion. This effect is demonstrated in Pinson et al. [12]. Whenever possible, we advocate the use of the partial matrix test design over a full matrix design. Each full matrix within the partial matrix test is associated with a different set of SRCs and HRCs. This adds much needed visual interest for the subject. HRCs that need to be directly compared should be put within the same matrix.

CHOOSING A SUBJECTIVE SCALE

The choice of subjective scale is surprisingly contentious. Each scale has strengths and weaknesses. Choose the scale that best matches your goal. There different scales for single stimulus (SS) and double stimulus (DS) experiments. In an SS test, the subject watches and rates each video sequence separately. In a DS test, the subject watches two or more versions of the same source video sequence during the rating process.

LISTING OF SUBJECTIVE SCALES

For the ACR from ITU-T Rec. P.910, the subject watches a video sequence and then is asked to rate it on a discrete, five-level scale (see Figure 2). Each level is associated with a word and a number: excellent = 5, good = 4, fair = 3, poor = 2, and bad = 1. Variations of the scale include nine levels, 11 levels, and a continuous scale with labels only at the end points. A continuous scale is a continuous line when presented to the subject and converted to a 100-level scale for the purposes of data analysis. Alternative labels may be needed for some experiments. ITU-T Rec. P.800, a speech quality subjective testing standard, provides alternate ACR wording examples for a listening-effort scale, and a loudness-preference scale. ACR is an SS method.

ITU-T Rec. P.910 identifies a variant, ACR with hidden reference (ACR-HR). In ACR-HR, each original is included in the experiment but not identified as such. The ratings for the originals are removed from the scores of the associated PVSs during data processing. High-quality originals are critical when using ACR-HR. If the quality of the original SRC drops from excellent (score 5) to fair (score 3), the available ACR-HR scale decreases from four to two units. This causes an inherent bias in the data, when comparing PVSs associated with two different SRCs.

The degradation category rating (DCR) method from P.910 also appeared in an old version of BT.500 under the name DS impairment scale (DSIS). DCR presents stimuli to subjects in pairs (see Figure 3). The original is presented first, and the subject is told that this is the original. The stimulus to be rated is presented second. The subject rates the difference in quality on a discrete, five-level, impairment scale: imperceptible = 5, perceptible but not annoying = 4, slightly annoying = 3, annoying = 2, and very annoying = 1. DCR is a DS method.

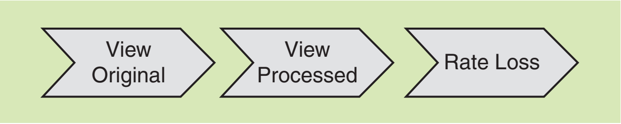
The pair comparison method from BT.500 also appears in P.800 under the name *comparison category rating* and in an old version of BT.500 under the name *double stimulus comparison scale* (*DSCS*). A pair of stimuli is presented to the subject; however, the order of stimuli is random (see Figure 4). If pair comparison is used to compare original and processed sequences (like DCR), then the original would be played first for approximately half of the trials, and the PVS would be played first for the rest of the trials. Pair comparison is the only method that can directly compare two different impaired versions of the same video sequence.

The subjects rate the quality of the second stimulus compared to the quality of the first on a discrete, seven-level scale: much better = 3, better = 2, slightly better = 1, about the same = 0, slightly worse = –1, worse = –2, and much worse = –3. Variations include a continuous scale (100 levels) and a discrete, two-level preference (better or worse).

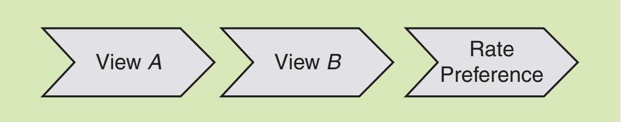
The double stimulus continuous quality scale (DSCQS) method from BT.500 involves four presentations of two stimuli, *A* and *B* (see Figure 5). One of these is the original, assigned randomly to position *A* or *B*. The subject is presented with stimulus *A*, then *B*, then *A* again, and then *B* again. Afterward, the subject rates *A* and *B* separately, each on a continuous scale showing the ACR labels (excellent, good, fair, poor, or bad).

The SS continuous quality evaluation (SSCQE) method from BT.500 presents the subject with a stimulus of long duration (e.g., 5–30 min). The subject has a slider that is constantly moved to reflect the subject's current opinion of the video quality (see Figure 6). Ratings are sampled every half second. SSCQE was intended for the analysis of monitoring applications and uses a continuous scale.

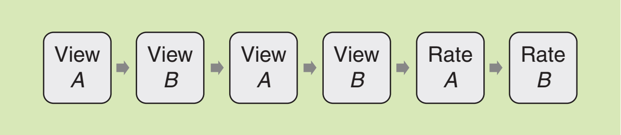
The SAMVIQ method from ITU-R Rec. BT.1788 uses a continuous scale, marked with the ACR labels. The test uses a computer interface, which presents the subject with multiple versions of the same SRC (see Figure 7). The subject may play each stimulus multiple times and may choose the order in which stimuli are rated. One of the stimuli is the original and explicitly labeled as such. Another stimulus is a hidden reference—identical to the original, but not labeled. The subject rates each version of one SRC and adjusts the ratings relative to each other.



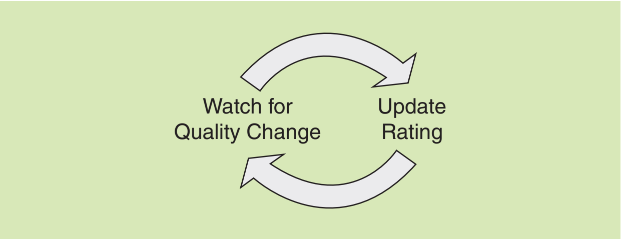
[FIG3] The DCR rating cycle: the subject watches the original video, then watches a processed version of that video, and finally rates the level of impairment.



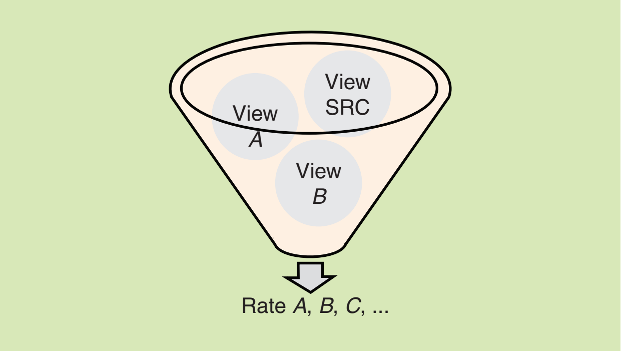
[FIG4] Pair comparison rating cycle: the subject watches video sequence *A*, then watches video sequence *B*, and rates *B* relative to *A*.



[FIG5] The DSCQS rating cycle: the subject watches video clip *A*, then *B*, *A* again, *B* again, and then rates *A* and *B*.



[FIG6] The SSCQE rating cycle: watch a long video sequence and continuously update a slider to reflect current opinion of the video quality.



[FIG7] The SAMVIQ rating cycle: the subject watches several versions of one SRC in any order. The ratings are adjusted until the subject is satisfied.

ANALYSIS OF SINGLE STIMULUS RATING METHODS

ACR with a five-level scale maximizes cognitive ease and the number of video sequences rated each minute [16]. ACR produces very repeatable subjective results, even across different groups of subjects, provided that the test design and instructions are carefully prepared [17]. Studies [16]–[18] compared ACR ratings with ratings gathered from DSCQS, DCR, and SAMVIQ. The rating scale choice had a minor impact on data accuracy.

The SS methods (ACR and SSCQE) have two weaknesses. The first is that some types of impairments are difficult to detect without an explicit comparison between two video sequences. For example, small color shifts are difficult to detect using ACR tests.

The second weakness is that SS ratings do not differentiate between impairments in the SRC and impairments intentionally added to the processed sequence. SS scores are often biased by subjects' opinions of the scene's aesthetics, composition, production quality, and subject matter, despite instructions to the contrary. ACR-HR offsets this flaw by removing the original video's rating during data analysis. Thus, ACR-HR ratings look more like DS ratings (i.e., a score of "5" on the ACR-HR scale means that the original video and PVS have identical quality). This technique only works better with high-quality SRCs.

SSCQE has strengths and weaknesses similar to ACR. SSCQE has the potential for allowing the most evaluations from a single subject in a short time, because there are no pauses between stimuli for ratings. Pinson and Wolf [19] demonstrated that SSCQE can be as accurate as DSCQS and pair comparison for rating short video sequences by using multiple randomizations, hidden reference removal, and the SSCQE score at the end of each sequence. A variety of devices have been used to implement an SSCQE slider, including a game station steering wheel [20] and a sensory glove [21]. True SSCQE data analysis is complex, as it requires time series analysis.

ANALYSIS OF DOUBLE STIMULUS RATING METHODS

DS methods address both problems of the SS method, at least to some extent. DS methods ask one of two basic questions.

The first DS question is "Which of these two sequences do you like better?" Pair comparison is the only DS method that directly answers this question. Inversion errors can occur but seem to be rare (e.g., the subject marks "the second sequence is much better" when they intended to mark "the second sequence is much worse"). Pair comparison takes approximately twice as long as ACR for the same number of stimuli.

Pair comparison is the obvious choice for detecting very small differences between two different impaired versions of the same video sequence. Doherty et al. [22] demonstrate the use of pair comparison to detect differences between frame rate conversion algorithms.

The second DS question is "How well does the impaired sequence reproduce the reference sequence?" DCR answers this question explicitly. DSCQS, SAMVIQ, and ACR-HR answer this question implicitly.

DCR takes approximately twice as long as ACR for the same number of stimuli. This is as fast as any DS method can claim, and inversion errors do not occur. Tominaga [16] concludes that DCR is more desirable than DSCQS or SAMVIQ, because of improved speed and ease of use, without loss of accuracy. DCR cannot be used to measure quality improvements—the rating scale does not allow a subject to say that the PVS is of higher quality than the source. DCR is the only method where subjects are unambiguously instructed to rate the perceptual difference between an original sequence and an impaired version of that sequence.

A quirk of DCR is that the original video will not be scored as perfect. That is, if the original video is played identically as both the "original" and "processed" video in Figure 3, the rating will be slightly lower than five. This imposes a systematic downward shift on all scores that bothers some researchers. (Pair comparison is the only subjective method that is likely to yield perfect scores for original video sequences.) This bias causes no problems for the data analysis.

DSCQS takes approximately four times as long as ACR for the same number of stimuli. The repeated viewings of stimuli A and B are intended to yield improved accuracy per subject for small quality differences, but this has not been proven.

Inversion errors are a problem for DSCQS. Inversion errors are impossible to detect and remove from the data, and this is perhaps the reason why DSCQS failed to show improved accuracy in [16] and [18]. It should be possible to avoid DSCQS inversion errors using an automated subjective testing system that swaps the order of the fourth and fifth steps in Figure 5.

SAMVIQ takes approximately twice as long as ACR for the same number of stimuli. SAMVIQ is slowed down by the ability of subjects to repeatedly play and compare sequences, yet sped up by presenting all versions of each SRC to the subject simultaneously. SAMVIQ is the only method that allows subjects to directly compare multiple versions of a single SRC.

The advantage of SAMVIQ is improved accuracy. SAMVIQ with 15 subjects is as precise as ACR with more than 22 subjects [17]. An open question is whether or not SAMVIQ's improved accuracy per subject yields an additional advantage for an experiment that focuses on a narrow range of quality.

DISCRETE VERSUS CONTINUOUS

Discrete levels are used for ACR, DCR, and pair comparison. Continuous scales are used for SSCQE and DSCQS. Researchers have explored continuous scales and different numbers of discrete levels for ACR and pair comparison. Tominaga et al. [16] showed that a five-level discrete scale provides a much easier cognitive task for the subject than an 11-level discrete scale or a continuous scale. Studies [16]–[18] demonstrate that the continuous scales do not improve measurement accuracy. This makes sense; research on human thought indicates that people can only hold about seven items in immediate memory [23]. (The Harvard Mind Brain Behavior Event Video Archive provides a nice summary of [23] at minute 8:30 of the video "The Cognitive Revolution at Fifty Plus or Minus One: A Conversation with Jerome Bruner, Susan Carey, Noam Chomsky, and George Miller—Part 1.")

When using a discrete scale, researchers disagree on whether or not the level numbers should be displayed to the subject. No consensus exists on this subject.

IMPLEMENTATION

PLAYING BROADCAST-QUALITY VIDEO ON A TELEVISION

The ideal video playback system plays uncompressed video flawlessly. Why use uncompressed video? With uncompressed video and perfect rendering come the guarantee that the playback system does not add new impairments to the video.

This ideal television playback/capture system costs about US$10,000 excluding the monitor, a cost that has been fairly consistent for the past decade. The components consist of a multiple core computer, a redundant array of independent disks (RAID), and a specialized board to play video from the RAID to the serial digital interface (SDI) high-definition SDI (HD-SDI) or high-definition multimedia interface (HDMI). The following companies currently produce professional grade video capture and playback cards that are compatible with professional editing suites: AJA, Bluefish444, Blackmagic, and Matrox. Bit-perfect playback and capture must be proven, which requires two systems (i.e., system 1 plays the video, system 2 captures it, and then a pixel-by-pixel comparison is performed). Common problems include insufficient RAID speed, operating system interruptions, antivirus software interruptions, and driver incompatibilities.

The alternative is to use dedicated hardware. There are too many professional-grade devices available to list in this article. Most of these devices compress the video slightly (e.g., four to ten times). Professional video devices ensure reliable video play and record capability, usually with no perceptual impairment.

PLAYING VIDEO ON A COMPUTER MONITOR

The ideal video playback system plays uncompressed, progressive video flawlessly from a computer hard drive to its monitor. If compressed playback is acceptable, the computer setup is simplified and the price drops. For some devices, only compressed video playback is easily available (e.g., smartphones), and reliable playback requires substantial compression (e.g., 30–250 times). Any added impairment from the coder, decoder, or display will confound the research data and may cause the data analysis to be misleading. To avoid this problem, identically compress all video sequences for the purposes of playback only. That is, after the videos have been impaired as specified in the experiment design, reencode all videos at the same (higher) bit rate for playback purposes. The goal is that any added playback impairments will be imposed identically on all videos.

When using a computer video playback system, you must calculate the appropriate level of compression yourself. The highest bit rate that guarantees flawless playback will minimize the perceptual impact of the recompression. To find this bit rate, encode a large variety of high-quality SRC, and play them to the target display repeatedly, while looking for playback problems (e.g., intermittent pauses and reduced frame rate).

THE TEST SESSIONS: AUTOMATED, EDITED, OR MANUAL

There are three options for playing video and recording scores during the actual subjective test: automated playback and scoring, edited sessions, and manual sessions.

When playing uncompressed or lightly compressed progressive video to a computer monitor, automated software provides an elegant solution for subjective testing. The software should identify subjects by ID number (see the section "Conducting the Experiment"), generate a unique randomized order of sequence presentation for each subject, implement the chosen method's rating cycle, ensure flawless playback for all subjects, present the rating scale after video playback finishes, record scores to a file, run each session separately, prompt for breaks, remove visual clutter from the screen, and either allow or disallow video replay. Three freely available software packages are AcrVQWin [24], Tally [25], and SubjectivePlayer [26].

If the subjects are allowed to replay the video, some subjects will, and this impacts the ratings. No consensus exists on the advisability of the replay option. Allowing replay provides the subject with an option other than guessing when their attention wandered—but conflicts arise with the usage paradigm for some video systems where rewind is not available. Thus, any article that describes a subjective test must specify whether or not videos could be replayed.

When subjective video quality testing began, the only option available was editing test tapes. Edited sessions work as well today as they did then, though the playback system is likely to be DVD, Blu-ray, three-dimensional (3-D) Blu-ray, or simply a long video file. No specialized software is required, and equipment costs are minimal. DVD or Blu-ray ensures consistent playback quality.

The concept is to edit together a long video sequence for each session. For example, when conducting an ACR test with 10-s sequences, the sequence would alternate between playing a 10-s sequence and playing 8 s of midlevel gray while the subject scores. This editing is simple yet prone to errors. A minimum of two different sequence orderings must be created to minimize the impact of ordering effects (i.e., the quality of clip N influences the perceived quality of clip $N+1$). Order effects can be reduced by randomizing the sessions (e.g., one subject sees session A, B, and C; another subject sees session C, B, and A). Ratings are entered either on a paper score sheet or on a small mobile device. Unlike an automated test, the ratings are not synchronized with the video playback. The audio track and text overlays keep the subject synchronized (e.g., please score clip 1). Subjects will sometimes make a mistake, and get off by one in their scores (e.g., record the quality of clip 9 where they were supposed to score clip 8). Data entry errors can occur when copying paper rating sheets into a spreadsheet.

The last option is a fully manual experiment. The experimenter can manually play each sequence in the desired order, ask the subject to choose a rating aloud, and record that rating themselves. This approach seems inelegant and the experimenter's behavior could influence ratings. It is, however, quite inexpensive and very practical.

We recommend playing midlevel gray between video sequences (i.e., Y = 128, Cb = 0, Cr = 0). The following MATLAB code will create this JPEG image, for video graphics array (480×640):

```
imwrite(zeros(480,640,'uint8') + 128,
        'Gray.jpg', 'jpg');
```

PRETEST

The purpose of the pretest is to check the experiment design for flaws. The pretest allows design problems to be fixed before too much time and money have been invested in the subjective test. Start by viewing the PVSs yourself. The resulting distribution of quality may show undesirable clusters. If you, the experiment designer, are unable to detect differences between most of your PVSs, the subjects will not either.

A pretest is often performed before some elements of the test are ready (e.g., no automation, no instructions, or an inappropriate environment). The pretest often includes only a subset of the scenes and impairments. A small, biased sampling of subjects (e.g., five to six coworkers or friends) is acceptable, because the goal is to look for design flaws. Example design flaws include problems subjects experienced during the experiment, a narrow range of quality (e.g., no statistically significant conclusions can be reached), and data bunching (e.g., many clips with nearly identical quality). Consider eliminating HRCs with nearly identical quality, when training or testing a model. Zieliński and Rumsey [27] identify potential sources of bias that should be considered during the pretest.

CONDUCTING THE EXPERIMENT

When administering the subjective test, the experimenter should not influence or bias any subject's behavior. To the extent possible, each subject's experience should be identical. The task of choosing an experiment administrator is very important. Good social skills and good communication skills are critical, as part of the administrator's job is to put the subjects at ease. The administrator must be guarded about the test itself and thus less likely to unintentionally influence the results. Questions that may influence the subject's behavior should be answered only after the subject's participation is over.

ETHICS AND INFORMED CONSENT

Awareness of ethical considerations in human testing arose from several infamous psychological and medical experiments. The Belmont Report [28], written in 1979 by the U.S. government, outlines the basic ethical principles in research involving human subjects. In 1991, the U.S. government published the Common Rule [29] for the protection of human subjects. While this policy applies only to U.S. Federal workers, it provides reasonable guidelines for ethical human testing and informed consent.

The first ethical consideration is privacy. Subjects' names must be kept private, and the researcher must ensure that the rating data cannot be used to identify subjects, even accidentally. The easiest and safest way to accomplish this is to identify subjects by number and to never record the number/name association. Second, subjects must be informed of potential risks. Video quality subjective experiments typically have no risk of benefit or harm. Third, subjects should be given an informed consent form to read and sign. The Common Rule contains guidance on appropriate information to include, such as a brief summary of the purpose of the experiment, the method used to keep people's names confidential, any risk or benefit to the subject, notification that participation is voluntary, and who to contact with questions about research subjects' rights or in the event of a research-related injury.

VISION TESTING

Vision testing is traditionally performed before the experiment begins. Unless you are an ophthalmologist, it is inappropriate to tell the subject whether or not they passed the vision test; all people should participate in the experiment, regardless of whether or not their data will be used.

ITU-R BT.500 and ITU-T P.910 require that subjects be screened for normal visual acuity (e.g., with glasses if worn) and normal color vision. Test the subject's distance vision using the Snellen eye chart at a range similar to that used during the experiment. Test color vision with the Ishihara Color Blindness plates under natural lighting (i.e., sunlight). The Ishihara plates should be replaced after about five years, because the colors fade, rendering the test inaccurate. These plates typically only test red–green color blindness, as that is the most common type.

There is some question as to whether there is a difference between ratings from people with normal vision and people who fail the distance vision or color vision test. Pinson et al. [15] found no significant difference in ratings; however, that was not the primary goal of the reported experiment. Moorthy et al. [30] present arguments against the use of vision tests. Regardless, these vision tests convey to the subject that they are participating in an important scientific experiment and should pay attention to their rating task.

SUBJECT DEMOGRAPHICS AND CONVENIENCE SAMPLING

The ideal in psychology is random sampling that perfectly matches the demographics of the population to be studied. This is most easily accomplished by outsourcing subject recruitment to a specialized company that performs market research through focus groups. The drawback is the high cost. Most video quality subjective tests use convenience sampling—i.e., a population of subjects that are easy to obtain. Universities tend to recruit students; large companies tend to recruit employees. To better represent the larger population, consider using a temporary hiring agency or online advertisements.

The problem with convenience sampling is that the research results may not generalize to the larger population. For video quality subjective testing, the relationships between variables will remain correct—but the MOS values will not be absolute (see Pinson et al. [15]). Be careful not to generalize your convenience sampling results into absolute thresholds (e.g., MPEG-2 at this bit rate will result in a quality of 4.0 or better).

INSTRUCTIONS, TRAINING SESSIONS, AND QUESTIONNAIRES

Instructions must be written out and agreed upon before the testing begins. All subjects must receive the same instructions. This eliminates one potential source of subject bias. The instructions should describe the rating cycle, the quality scale, how to record ratings, quirks of your playback system or environment, behavior to be avoided, and the scenario (e.g., watching free video clips on a mobile device and watching a pay-per-view movie). The instructions should include: "Please do not base your opinion on the content of the scene or the quality of the acting." Still, ratings inevitably include both the clip's artistic quality and its technical quality. This is why subjective tests normally include the original video for comparison. After presenting the instructions, ask the subjects if they have any questions.

The training session immediately follows the instructions. The training session serves two purposes. The first is to demonstrate the task. This is easily accomplished with two or three rating cycles. The second purpose is to familiarize the subject with the range of quality and type of impairments in the experiment. This may take much longer (e.g., 5–20 sequences). The SRC used for training session should not appear in the rest of the experiment.

A questionnaire can be used after the rating sessions, to gather additional information. Questionnaires can potentially provide feedback on problems the subject had with the test, whether or not the subject understood the task, and whether the subject noticed a problem with your test setup. A written questionnaire is preferable to asking questions aloud because people are more likely to be blunt and provide additional information. Questionnaires can also be used to understand QoE (see the section "Data Analysis Techniques"). No standard questionnaire exists today.

WRITING THE REPORT

The ultimate goal of a video quality subjective test inevitably includes publishing the results, either internally or externally. The report of results should fully describe

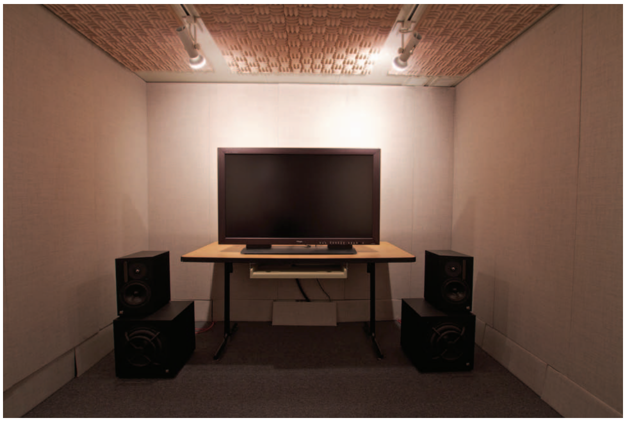
- the goal of the experiment
- the environment (e.g., monitor brand and model, lighting level in lux, viewing distance in screen heights, and picture of environment)
- the test methodology (e.g., ITU recommendation and any departures)
- the rating method
- the SRC (e.g., quantity and sample frames)
- the HRC (e.g., quantity, coding algorithm, bit rate, and transmission error level)
- the experiment design (e.g., full matrix or partial matrix or other, number of PVS)
- the test sessions (e.g., number, duration, playback mechanism, playback compression characteristics, and software used to control the test)
- the subjects (e.g., number of, age and gender distribution)
- the mechanism used to obtain subjects
- the data analysis results.

Always include a picture of the viewing environment in the report. This will provide readers with an improved understanding of the environment (see Figure 8). Brunnström et al. [31] is an example of a superior experiment report.

For privacy reasons, the names of subjects should not be mentioned in any report. Care should be taken when explicitly mentioning vendor names. If the experiment was not designed to directly compare the quality of those vendors' equipment, a comparative analysis might be biased. In such cases, the vendor names should be omitted from external reports.

DATA ANALYSIS TECHNIQUES

Any data analysis is divided into three specific steps: clean the data, choose the correct analytical technique (this step should be done before the subjective study is run), and interpret the results. Each step's method has to fit the problem under investigation. To begin, key statistical concepts will be described.



[FIG8] A picture of the viewing environment should be included in the final report. This sample environment shows a sound isolation booth. A Blu-ray player outside the room plays audiovisual sequences on a broadcast-quality monitor and speakers.

MEASUREMENT SCALES, AGREEMENT, AND ASSOCIATION

The values related to a particular variable (e.g., bit rate) can be measured in different ways [32]. Measurement scales are classified and divided into four types: ratio, interval, ordinal, and nominal. For each type of measurement scale, correct statistical techniques exist and should be used. The ratio measurement scale makes it possible to define a distance between any two values and compute their ratio. For example, the distance between 3 Mbit/s and 6 Mbit/s is 3 Mbit/s, and the second bit rate is two times larger than the first. With the interval measurement scale, the distance between each point is the same but the measured numbers are arbitrary. For example, consider encoder bit rate setup categorized to three values $1 = 3$ Mbit/s, $2 = 5$ Mbit/s, and $3 = 7$ Mbit/s). The measured values 1, 2, and 3 cannot be compared using a ratio (i.e., 3 does not have bit rate three times as high as 1), but the distance between 1 and 2 is the same as the distance between 2 and 3. With an ordinal measurement scale, an order of values can be found but exact distances cannot. For example, bit rate category "medium" has a higher bit rate than "low" and lower than "high." Nevertheless, with an ordinal scale, we cannot determine whether a value "low" has the same distance to "medium" as "medium" to "high." With nominal, each value is different and an order cannot be determined. For example, encoders *A*, *B*, and *C* cannot be ordered without focusing on a specific feature, like price or encoding speed.

Agreement means that subjects should give the same quality ratings, and we only tolerate differences that are caused by a random distribution of measurement noise. On the other hand, association requires only that subjects follow the same pattern. So, if two subjects agree, they also associate, but the opposite is not true [33]. For example, if subject one is always scoring one point lower than subject two (where possible), they do not agree at all but they do associate perfectly. For subjective experiments focused on quality, agreement is not expected, since a subject can be more or less discriminating. On the other hand, if there is no association it can be explained only if a subject is more tolerant than average for some impairments and less for others. This is not impossible, but an experimenter may choose to consider a subject who does not associate with others to be irrelevant. Perhaps the rating task was too difficult or was not understood properly.

CLEANING THE DATA

Before using any statistical technique, the data has to be cleaned. This involves first detecting irrelevant subjects and second detecting errors in the experiment setup. Detecting an irrelevant subject depends on the way the test was run. It is usual to monitor a subject in a lab; in this case, we can be sure a subject did the test. If this does not happen, the first screening technique is focused on finding out whether a subject actually did the test. Gardlo et al. [34] describe some techniques such as adding content questions, e.g., "Was a car present in the scene?" Such questions reveal whether a subject actually saw a particular sequence. If a test is run in a controlled environment, detecting whether a subject did the test is usually trivial. Nevertheless, the subject could misunderstand the test, the test could be too difficult, or the subject might not pay attention. Therefore, screening is still a necessary and important step.

The screening technique described in the ITU-R Rec. BT-500 is based on the subject agreement and measuring scale being ratios. The basic concept is to measure how often a subject's answers do not fit the confidence interval created by the other subjects' answers. (A MATLAB code for performing this test can be found at http://www.its.bldrdoc.gov/resources/video-quality-research/guides-and-tutorials/subject-screening-overview.aspx.) Subjects do not have to agree among themselves, so we will focus on techniques based on association rather than agreement.

The most popular way to measure association is the Pearson correlation. This technique was used to screen subjects in the Video Quality Experts Group (VQEG) HDTV validation test plan [35]. The Pearson correlation is based on the assumption that the measurement scale is a ratio, and Pearson correlation interpretation and tests generally assume the data have a normal distribution. If the quality scale is short (e.g., ACR), the Pearson correlation should be changed to the Spearman correlation, which is based on the weaker assumption that the measurement scale is ordinal. For a short scale, it is difficult to assume that the answer distribution is close to normal (especially if most answers are close to one of the scale borders) and that the distances between answers are the same. Computing the Pearson or Spearman correlation requires two vectors: u_i (a single subject's ratings) and $\overline{u}_i$ (average of the other subjects' ratings) for all sequences. The MATLAB Statistics Toolbox functions are as follows:

```
 a=corr (ui, udashi, 'Type', 'Pearson');
b=corr (ui, udashi, 'Type', 'Spearman');
```

where u_i is denoted by `ui` and $\overline{u}_i$ is denoted by `udashi`.

The question is: At what threshold should a subject be discarded? A correlation that is statistically greater than zero does not guarantee that a subject is relevant. A good example occurred during a VQEG HDTV test [35], the goal of which was to choose the best objective metric for HDTV. For an experiment that used edited sessions on Blu-ray discs, the scoring time was too short for one subject, so he did not see the first few seconds of some sequences. The Pearson correlation was statistically significantly higher than zero (0.633) but much lower than for other subjects' correlation (the lowest being 0.784 and the average 0.877). The VQEG rule is that a subject should be discarded if the correlation is below 0.75 for television (TV) and mobile applications. The disadvantage of using correlation for data cleaning is that the correct threshold must be found experimentally, based on how difficult it is for subjects to evaluate a particular service. The 0.75 threshold cannot simply be used for new services (e.g., 3-D TV, ultraHD) or a different association metric (e.g., Spearman correlation).

These two methodologies (i.e., as described in Rec BT.500 and based on Pearson correlation analysis) are by far the most commonly used to detect irrelevant subjects. They rely on the ratio measuring scale. Since subjective experiments are often performed on a short measuring scale, which is ordinal rather than ratio, some different methods are needed. Adejumo et al. [33] and Gibbons [36] provide descriptions of numerous different agreement and association metrics. A commonly used agreement metric using an ordinal scale is named *Cohen's Kappa coefficient*. It is widely used by the psychology community when subject agreement is especially important. The Kappa coefficient can be computed using the MATLAB function created by Cardillo [37]. No more details are presented in this article since association rather than agreement metrics should be used in case of quality tests. Nevertheless, the interested reader can find details in [33].

Kendall's tau can be an alternative solution when an association metric is needed and the measuring scale is ordinal. Kendall's tau is the difference between the probability of concordance π_c and discordance π_d, given by: $\tau = \pi_c - \pi_d$. Kendall's tau is an easy to interpret parameter, since it refers to concordance and discordance between two vectors. Gibbons [36] provides more details on when Kendall's tau should be used. The MATLAB function is

```
a=corr (ui, udashi, 'Type', 'Kendall');
```

A subject can have a low association with other subjects because the subject did not pay attention to the task or because an error occurred in the experiment setup. While an experiment is being run, many different problems can occur. An intermittent video playback problem in the test interface can cause occasional added visual impairment (e.g., freezes during a sequence that should have played smoothly). Video clips can be played in the incorrect order, so that the written video clip/rating association is incorrect (e.g., an editing error in a prerecorded sequence, a coding bug in automated playback software). Some sequences can be different from the description (e.g., encoded at the wrong bit rate). A time synchronization error can occur between subject ratings and videos (e.g., a subject scoring on paper rating sheets scores clip *N* in the box for clip *N*+1). All of these problems have been observed by the authors of this article.

Specific problems call for specific solutions, so it is impossible to describe a general procedure. Nevertheless, our experience shows that the following three steps are very helpful. First, if prerecorded sequences are used, the results obtained for a PVS should be statistically the same when it appears in different recordings. Second, the quality ratings can be compared to the experimenter's expectations, and sequences with large differences should be checked. Third, the association between users within sessions should be checked. If a user is well associated with others for one session and poorly for another, this indicates a problem with that session.

The most difficult problems to detect are rare interface failures resulting in degradation of the watched video (e.g., the automated test software pauses during video playback for 2% of the sequences). This problem is best detected during the pretest (e.g., the experimenter takes the pretest, and then compares his or her scores to the expected quality). Intermittent playback failures will cause all subjects to score atypically low quality for a few, random clips. One way to find such inconsistencies is to compare a user's association with other users for all sequences with that user's association with other users for a subset of sequences. If the percentage of errors in a subset is higher than in the whole set it is easier to detect them.

If the data cleaning eliminates more than two or three subjects, something may be wrong with your test procedure. To put this into perspective, only one of the 214 subjects in [15] was eliminated for being irrelevant (see [38] to view the individual subject ratings).

DATA ANALYSIS

After screening subjects and ensuring that all of them performed the experiment properly, the final and most important analysis can be run. The section "Goal of Experiment and Design Consequences" described the different reasons to run subjective experiments. These different reasons call for different data analyses.

ANSWERING A QUESTION

An experiment designed to answer a question contains different conditions, which are most often different HRCs but could also be different subjects or different SRCs. Different conditions generate groups of results that can be compared to answer specific question. Therefore, answering a question can be reduced to comparing subsets of subjective experiment results. The most common way to compare two groups is to answer the question of whether the results are statistically the same or not. This question will be answered with a specific significance level. Most often, 5% significance (i.e., 95% confidence) is chosen. This is the default in many MATLAB statistical tests and is specified as 0.05.

The most popular technique for comparing two groups is the Student's t-test. The goal of the Student's t-test is to validate if the difference between the mean values of two groups has a particular value. For example, if our goal is to validate whether the quality obtained for HRC 1 and HRC 2 are the same we should compare vectors $\overline{u}_1$ and $\overline{u}_2$. Each element of the $\overline{u}_i$ vector is a value obtained for the same HRC with different other conditions (e.g., SRCs or repetitions). It is very important to have the same order of conditions in both vectors. The MATLAB function for computing the Student's t-test is: `[h, p]=ttest(u1-u2);` where `h` is one if the difference is statistically different from zero and zero if it is not, and `p` is the corresponding *p*-value, which has to be larger than 0.05 to conclude that the obtained difference is not statistically significant.

If the goal is to compare multiple groups, then the methodology and significance level must be adjusted to maintain the same significance level for a group as for single comparison. The commonly used methodology for comparing multiple groups is one-way analysis of variance (ANOVA). The MATLAB `anova1` computes one-way ANOVA. A handy way to call this function is to specify a vector of all compared values (e.g., MOSs) and a vector of tags describing groups. For example, `u= [2.3,3.2,1.2,2.4,3.2]` and `g= ['A','B','B','A','C']` means that the first and fourth values belong to group A, the second and third to group B and the last one to group C. `anova1` is called using: `p = anova1(u,g);` where `u` is the compared values vector, `g` is the grouping vector, and output `p` is the *p*-value. Similarly to `ttest`, a value of `p` smaller than 0.05 indicates that at least one group is different.

The disadvantage of both the Student's t-test and ANOVA is the assumption that the data come from a normal distribution, i.e., they follow a specific distribution and can be measured on a ratio or interval scale. This can be validated by the Kolmogorow–Smirnow (small sample) or chi square (large sample) test. If one of those assumptions is not met, different statistical methods should be used. The Student's t-test should be changed to the Mann–Whitney U-test and the one-way ANOVA should be changed to the Kruskal–Wallis test. The Mann-Whitney U-test compares medians not means, and as such it needs only an ordinal measuring scale. The equivalent of this test in MATLAB is `[p, h] = ranksum(u1, u2);` where the `p, h, u1,` and `u2` parameters are the same as for the `ttest` function but ordered differently. The multiple-group comparison version of the Mann–Whitney U-test is the Kruskal–Wallis test, which can be called similarly to the `anova1` function: `p = kruskalwallis(u,g);`

The discrimination powers of the Student's t-test and the ANOVA test are greater than those of the Mann–Whitney U-test and the Kruskal–Wallis test. The cost for this increased discrimination is the requirement of a normal distribution and interval measuring scale. Choosing the Mann–Whitney U-test or Kruskal–Wallis test leads to more conservative conclusions. If a test with the weaker assumption shows differences between groups, then the Student's t-test and ANOVA will show it as well. The opposite is not necessarily true.

A useful MATLAB tool for comparing multiple groups is the `multcompare` function. Both `anova1` and `kruskalwallis` functions can return more than one parameter. After three parameters are returned, `multcompare` can be run by using

```
[p,table,stats] = kruskalwallis(u,g);
multcompare(stats);
```

This function generates a handy interactive plot that makes it easy to compare groups.

TRAINING A METRIC OR ALGORITHM

Subjective experiments maximize measurement accuracy, but also increase cost and time taken. They cannot be used to monitor a service. Therefore, it is a common practice to build a metric that objectively emulates a video quality (i.e., MOS). When training a metric or algorithm, the goal is to find a function that links explanatory variables to a dependent variable. Example explanatory variables are bit rate, packet loss ratio, quality estimation parameters extracted from the video, and subject age. The dependent variable is most often video quality.

The easiest solution is to design a linear model using linear regression. A linear model is a linear function of model parameters, but not necessary a linear function of explanatory variables. Nonlinearities in the explanatory variables are detected and removed (e.g., using square or square root functions), and it is common to use interactions between explanatory variables (e.g., the product of two explanatory variables). According to the above description an example linear model is given by the equation $u = a_0 + a_1 b + a_2 \log ta + a_3 b \log ta$, where u is estimated MOS, a_i are model parameters, b is a bit rate, and ta is the temporal activity of the SRC. A linear model can be estimated in MATLAB by the `glmfit` function, which returns both the estimated values and the p-values of each estimated parameter.

While training the linear model, the researcher examines and understands the relationship between the candidate explanatory variables and the dependent variable. Example techniques include examining the ability of a single explanatory variable to predict the dependent variable (e.g., using the Pearson correlation or root mean square error), and plotting the explanatory variable against the dependent variable to find nonlinearities or outliers. Fox [39] provides instruction on techniques for applying linear regression analysis. The advantage of linear regression is that the resulting linear model is typically easy to explain and understand.

Alternatives to classical linear regression are methods based on machine learning. Many techniques are available. In this article, only three are mentioned. Genetic programming-based symbolic regression analyzes a large number of different models, thus helping to build a model that is similar to a linear model [40]. The advantage of this technique is that the output is easy to interpret. Partial least squares regression is more difficult to interpret but has the advantage of optimizing explanatory variables. Because of the use of principal component analysis, the final output is as simple as possible for a given prediction accuracy using explanatory variables that contain the most significant information [41]. Random neural networks are even more difficult to interpret but can approximate different nonlinear functions [42].

Machine-learning algorithms must be used with care to not over train the model. A typical machine-learning model contains lots of parameters, and relatively little subjective data are typically available to train a video quality metric.

All of the previously presented solutions model MOS (i.e., the average of many ratings), not the actual subjective ratings. If quality is measured on a scale with a small number of levels, each rating level's probability can be predicted using the generalized linear model (GLM). GLM is able to model multinomial distribution. A detail description of using GLM is given in [43].

ANALYZING A METRIC OR ALGORITHM

To analyze a metric, its predictions and subjective results have to be compared. The algorithm that fits the subjective data best should be chosen. This analysis has to address two specific realities of subjective experiments. Previous research shows that two instances of the same subjective experiment repeated in two different laboratories can have high association (measured by correlation) but the results are not identical [3], [15]. Since the results of the two subjective experiments results have high association but not necessarily agreement, a metric should associate with the validation subjective experiment but it does not have to agree (i.e., an offset is possible).

The final conclusion is that metrics should be validated by association rather than by agreement, or agreement should be measured for the metric after the values have been transferred to a common scale. In addition, two metrics can differ due to randomness related to the subjective experiment. Such metrics should be called "the same" even if the agreement or association metric is superior for one of them. The methodology for addressing the problems described above is used by VQEG and is described in ITU-T Rec. P.1401.

THOUGHTS ON QUALITY OF EXPERIENCE

We have provided detailed information for conducting video quality subjective tests. Video quality is one aspect of a larger topic—QoE. Compared to video quality testing, QoE testing is in its infancy, and no step-by-step tutorial is available at this time. Instead, this final section summarizes some QoE definitions and frameworks. This overview points out limitations of video quality subjective tests and identifies areas where QoE issues impact experiment design, monitor selection, subject demographics, post-test questionnaires and, as a consequence of these choices, the strength of the conclusions that can be reached.

Video quality is just one aspect of QoE. According to Le Callet et al. [44], QoE is the degree of delight or annoyance the user receives from an application or service. It results from the fulfillment of the user's expectations (in light of his or her personality and current state) with respect to the utility and/or enjoyment of the application or service. More succinctly, QoE is a measure of how well a service or an application meets the user's expectation of quality (EoQ) [45]. Different artifacts arising along an end-to-end service delivery chain may result in QoE that does not meet a user's EoQ. However, each service provider is expected to aim at the condition QoE = EoQ [46], ensuring revenues, reducing churn, and increasing customer satisfaction.

Today, each of us is a consumer of multimedia services and knows how many variables influence our EoQ. Therefore, a holistic QoE approach should span the whole telecommunication ecosystem combining user behavior, technical issues, and business models as proposed in [44] and [47].

Batteram et al. [45] propose three dimensions that can be used to express QoE: service availability, service responsiveness, and media quality (i.e., audio and video quality). Service availability is a measure of whether the user can use the desired service, while responsiveness is the time to get the service answer. Media quality relates to all artifacts generated by compression and packet network delivery that deteriorate the user's perception. Audio and video quality subjective tests measure media quality, but fail to quantify the impact of service availability and responsiveness.

On the other hand, Marez and Moor [46] point out that QoE may depend on many service context-of-use factors (i.e., the actual conditions under which an application is used). The service has to be paid for through some provider-defined business mode (e.g., transaction, subscription, and advertisement). The underlying network technology (e.g., wired, wireless, and satellite) impacts QoE, as do other technological factors. Personal, social, cultural, and education issues are influential, and a user's EoQ is modified by the location or device used for service consumption. There is interest in extending video quality subjective testing techniques but as of yet no established solution (e.g., a way to measure MOS that accounts for screen size differences).

The multivariate structure of QoE may suggest initially that, from the QoE analytical modeling point of view, numerous analysis models could be deployed to understand the relationships between variables and their relevance to the actual QoE problem being studied [44]. An expert panel [46] found about 60 multidisciplinary methods (both qualitative and quantitative) suited for QoE investigations.

While the rating scales in the section "Ethics and Informed Consent" are intervals (which define the ratio, interval, ordinal, and nominal measurement scales), QoE variables are often ordinal (e.g., satisfied, neutral, and dissatisfied) or nominal (e.g., gender, user profile, device, and content type). These category variables differ radically from interval variables because distances between categories are not defined and subjects can interpret the categories differently. Thus, most of the techniques from the section "Data Analysis Techniques" are inappropriate. The proper tool for dealing with such unmeasurable variables is categorical data analysis (i.e., multicategory logit models). These techniques are more complicated, and the results derived are a bit more difficult to interpret.

Customer satisfaction surveys solve this problem by using a variety of latent trait models (LTMs). For example, the Item Rasch Theory is the simplest LTM model. The LTM is a powerful approach as it can relate manifest variables (i.e., service features that can be readily judged by a tester) with latent traits (i.e., the tester's experience with a service)—provided that the questionnaire is properly designed. The LTM approach was recently suggested by [48] and [49] as a proper tool for 3-D video quality analysis.

In summary, QoE uses multiple dimensions to measure different users' experiences of service received and relate their experiences to parameters of a service delivery chain and a service context-of-use. A reliable QoE measurement calls for a multidisciplinary approach (e.g., operations research, customer satisfaction surveys, and sociology), because of the different nature of the variates involved. Users' experiences may differ even if they use the application in the same context and under the same network conditions. Therefore, to arrive at a valid QoE assessment, it is necessary to conduct tests with large numbers of subjects. Although such subjective tests are time and resource consuming, emerging crowdsource QoE assessment has recently appeared as a solution [50]. QoE is gaining increasing momentum among researchers, service providers, and network operators; this may eventually result in the implementation of user-centric services.

ACKNOWLEDGMENTS

We would like to thank the many researchers whose questions or advice contributed to this article, as well as Florence Agboma and Jessica Hooker. The work of Lucjan Janowski and Zdzislaw Papir was supported by the AGH University of Science and Technology under contract 15.11.230.075.

AUTHORS

Margaret Pinson (margaret@its.bldrdoc.gov) received her B.S. and M.S. degrees in computer science from the University of Colorado at Boulder in 1988 and 1990. Since 1988, she has been investigating improved methods for assessing video quality at the Institute for Telecommunication Sciences, an office of the National Telecommunication and Information Administration, in Boulder, Colorado. She is a cochair of VQEG and a cochair of the Independent Lab Group in the VQEG, an associate rapporteur of Question 12 in the ITU-T Study Group 9, and the administrator of the Consumer Digital Video Library.

Lucjan Janowski (janowski@kt.agh.edu.pl) is an assistant professor with the Department of Telecommunications, AGH University of Science and Technology. He received his Ph.D. degree in telecommunications in 2006 from the AGH. In 2007, he worked in a postdoctoral position at the Centre National de la Recherche Scientifique (CNRS), LAAS (Laboratory for Analysis and Architecture of Systems of CNRS) in France, where he prepared both malicious traffic analysis and anomaly detection algorithms. In 2010–2011, he spent half a year in a postdoctoral position at the University of Geneva, working on quality of experience (QoE) for health applications. His main interests are statistics and probabilistic modeling of subjects and subjective rates used in QoE evaluation.

Zdzisław Papir (papir@kt.agh.edu.pl) is a professor and a deputy chair with the Department of Telecommunications, AGH University of Science and Technology in Cracow, Poland. From 1994 to 1995 he served as a network design department manager for Polish Cable Television. He was a guest coeditor for *IEEE Communications Magazine* from 1999 to 2006, responsible for the broadband access series. He participates in several R&D projects under the Information Society Technologies Work Program of the European Commission, responsible for network performance evaluation and quality assessment of communication services. He is also an information and communication technology expert as appointed by the European Commission.

REFERENCES

[1] (2014, Feb. 5). Cisco visual networking index: Global mobile data traffic forecast update, 2013–2018. [Online]. Available: http://www.cisco.com/

[2] M. Kuniavsky, A. Moed, and E. Goodman, *Observing the User Experience: A Practitioner's Guide to User Research*. Waltham, MA: Elsevier Science, 2012.

[3] G. Cermak et al. (2008, Mar. 28). Final report from the video quality experts group on the validation of objective models of multimedia quality assessment, phase I. [Online]. Available: https://www.vqeg.org

[4] A. Younkin and P. Corriveau, "The effects of quantity and location of video playback errors on the average end-user's experience," in *Proc. IEEE Int. Symp. Broadband Multimedia Systems and Broadcasting*, 2008, pp. 1–5.

[5] M. H. Pinson, S. Wolf, and G. Cermak, "HDTV subjective quality of H.264 vs. MPEG-2, with and without Packet Loss," *IEEE Trans. Broadcast*., vol. 56, no. 1, pp. 86–91, Mar. 2010.

[6] K. Brunnström, L. Nordtström, and B. Andreén, "Visual experience of quality degradations when viewing computer and notebook displays from an oblique angle," *J. Soc. Inform. Display*, vol. 19, no. 5, pp. 387–397, May 2011.

[7] S. Voran and S. Wolf, "The development and evaluation of an objective video quality assessment system that emulates human viewing panels," in *Proc. Int. Broadcasting Convention (IBC)*, July 1992, pp. 504–508.

[8] Q. Huynh-Thu and M. Ghanbari, "Temporal aspect of perceived quality in mobile video broadcasting," *IEEE Trans. Broadcast*., vol. 54, no. 3, pp. 641–651, Sept. 2008.

[9] S. Voran and A. Catellier, "When should speech coding quality increases be allowed in talk-spurts?" in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, May 2013, pp. 8149–8153.

[10] "Subjective quality evaluation of audio and audiovisual multiparty telemeetings," ITU-T Rec. P.1301, 2012.

[11] G. Berndtsson, M. Folkesson, and V. Kulyk, "Subjective quality assessment of video conferences and telemeetings," in *Proc. Packet Video Workshop (PV)*, May 2012, pp. 25–30.

[12] M. H. Pinson, M. Barkowsky, and P. Le Callet, "Selecting scenes for 2D and 3D subjective video quality tests," *EURASIP J. Image Video Processing*, vol. 2013, no. 1.

[13] Y. Niu and F. Liu, "What makes a professional video? A computational aesthetics approach," *IEEE Trans. Circuits Syst. Video Technol*., vol. 22, no. 7, July 2012, pp. 1037–1049.

[14] A. Catellier, M. Pinson, W. Ingram, and A. Webster, "Impact of mobile devices and usage location on perceived multimedia quality," in *Proc. Int. Workshop on Quality of Multimedia Experience* (QoMEX), July 2012, pp. 39–44.

[15] M. H. Pinson, L. Janowski, R. Pépion, Q. Huynh-Thu, Ch. Schmidmer, P. Corriveau, A. Younkin, P. Le Callet, M. Barkowsky, and W. Ingram, "The influence of subjects and environment on audiovisual subjective test: an international study," *IEEE J. Select. Topics Signal Processing*, vol. 6, no. 6, pp. 640–651, Oct. 2012.

[16] T. Tominaga, T. Hayashi, J. Okamoto, and A. Takahashi, "Performance comparisons of subjective quality assessment methods for mobile video," in *Proc. Int. Workshop on Quality of Multimedia Experience* (QoMEX), June 2010, pp. 82–87.

[17] Q. Huynh-Thu and M. Ghanbari, "A comparison of subjective video quality assessment methods for low-bit rate and low-resolution video," in *Proc. Signal and Image Processing*, M. W. Marcellin, Ed., Honolulu, Hawaii, USA, 2005, vol. 479, pp. 70–76.

[18] M. D. Brotherton, Q. Huynh-Thu, D. S. Hands, and K. Brunnström, "Subjective multimedia quality assessment," *IEICE Trans. Fundamentals, Electron. Commun. Comput. Sci*., vol. E89-A, no. 11, pp. 2920–2932, 2006.

[19] M. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," in *Proc. SPIE Video Communications and Image Processing Conf*., Lugano, Switzerland, July 2003, pp. 8–11.

[20] T. Liu, G. Cash, N. Narvekar, and J. Bloom, "Continuous mobile video subjective quality assessment using gaming steering wheel," in *Proc. Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, Jan. 2012.

[21] S. Buchinger, W. Robitza, M. Nezveda, M. Sack, P. Hummelbrunner, and H. Hlavacs, "Slider or glove? Proposing an alternative quality rating methodology," in *Proc. 6th Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, Jan. 2010.

[22] R. A. Doherty, A. C. Younkin, and P. J. Corriveau, "Paired comparison analysis for frame rate conversion algorithms," in *Proc. Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, Jan. 2009.

[23] G. A. Miller. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychol. Rev*. [Online]. 63, pp. 81–97. Available: http://www.musanim.com/miller1956/

[24] J. Jonsson and K. Brunnström. (2007). *Getting Started with AcrVQWin*. [Online]. Acreo AB, Kista: Sweden. Available: https://www.acreo.se/acrvqwin

[25] A. Jain, C. Bal, and T. Nguyen, "TALLY: A web-based subjective testing tool," in *Proc. Int. Workshop on Quality of Multimedia Experience* (QoMEX), July 2013, pp. 128–129.

[26] S. Buchinger, W. Robitza, M. Nezveda, P. Hummelbrunner, and H. Hlavacs. (2010, Dec. 15). Towards a comparable and reproducible subjective outdoor multimedia quality assessment, *Third Euro-NF IA.7.5 Workshop on Socio-Economic Issues of Networks of the Future*. [Online]. Available: http://code.google.com/p/subjectiveplayer/

[27] S. Zieliński and F. Rumsey, "On some biases encountered in modern audio quality listening tests—A review," *J. Audio Eng. Soc*., vol. 56, no. 6, June 2008, pp. 427–451.

[28] U.S. Department of Health & Human Services. (1979, Apr. 18). Ethical principles and guidelines for the protection of human subjects of research. [Online]. Available: http://www.hhs.gov

[29] U.S. Department of Health & Human Services. (1991). Federal policy for the protection of human subjects ('Common Rule'), 45 CFR part 46 [Online]. Available: http://www.hhs.gov

[30] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. de Veciana, "Video quality assessment on mobile devices: subjective, behavioral and objective studies," *IEEE J. Select. Topics Signal Processing,* vol. 6, no. 6, pp. 652–671, Oct. 2012.

[31] K. Brunnström, L. Nordström, and B. Andrén, "Visual experience of quality degradation when viewing computer and notebooks displays from an oblique angle," *J. Soc. Inform. Display*, vol. 19, no. 5, pp. 387–397, Dec. 2011.

[32] A. Agresti, *Categorical Data Analysis*, 2nd ed. Hoboken, NJ: Wiley, 2002.

[33] A. O. Adejumo, C. Heumann, and H. Toutenburg, "A review of agreement measure as a subset of association measure between raters," Sonderforschungsbereich 386 der Ludwig-Maximilians-Universität München, no. 385, 2004.

[34] B. Gardlo, M. Ries, and T. Hossfeld, "Impact of screening technique on crowdsourcing QoE assessments," Radioelektronika (RADIOELEKTRONIKA), in *Proc. 2012 22nd Int. Conf*., 17–18 Apr. 2012, pp. 1–4.

[35] M. Pinson et al. (2010). Report on the validation of video quality models for high definition video content, in *VQEG*. [Online]. Available: https://www.vqeg.org

[36] J. D. Gibbons, *Nonparametric Measures of Association*. Thousand Oaks, CA: SAGE, 1993.

[37] G. Cardillo. (2009, Dec.). Cohen's kappa: Compute the Cohen's kappa ratio on a 2x2 matrix [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/15365

[38] M. H. Pinson, L. Janowski, R. Pépion, Q. Huynh-Thu, Ch. Schmidmer, P. Corriveau, A. Younkin, P. Le Callet, M. Barkowsky, and W. Ingram, "Subjective and objective evaluation of an audiovisual subjective dataset for research and development," in *Proc. Int. Workshop on Quality of Multimedia Experience (QoMEX)*, July 2013, pp. 30–31.

[39] J. Fox, *Applied Regression Analysis and Generalized Linear Models*, Thousand Oaks, CA: Sage, 2008.

[40] N. Staelens, D. Deschrijver, E. Vladislavleva, B. Vermeulen, T. Dhaene, and P. Demeester, "Constructing a no-reference H.264/AVC bitstream-based video quality metric using genetic programming-based symbolic regression," *IEEE Trans. Circuits Syst. Video Technol*., vol. 23, no. 8, pp. 1322–1333, Aug. 2013.

[41] C. Keimel, J. Habigt, M. Klimpke, and K. Diepold, "Design of no-reference video quality metrics with multiway partial least squares regression," in *Proc. Int. Workshop on Quality of Multimedia Experience (QoMEX)*, 7–9 Sept. 2011, pp. 49, 54.

[42] K. D. Singh and G. Rubino, "No-reference quality of Experience monitoring in DVB-H networks," in *Proc. Wireless Telecommunications Symp. (WTS)*, 21–23 Apr. 2010, pp. 1–6.

[43] L. Janowski and Z. Papir, "Modeling subjective tests of quality of experience with a generalized linear model," in *Proc. Int. Workshop on Quality of Multimedia Experience (QoMEX)*, 29–31 July 2009, pp. 35–40.

[44] A. Raake and S. Egger, "Quality and quality of experience," in *Quality of Experience: Advanced Concepts, Applications and Methods*, by S. Möller and A. Raake, Ed. New York: Springer, 2014, ch. 2, pp. 11–33.

[45] H. Batteram, G. Damm, A. Mukhopadhyay, L. Philippart, R. Odysseos, and C. Urrutia-Valdés, "Delivering quality of experience in multimedia networks," *Bell Labs Tech. J*., vol. 15, no. 1, pp. 175–194, 2010.

[46] L. De Marez and K. De Moor, "The challenge of user- and QoE-centric research and product development in today's ICT-environment," *Observ. J*., vol. 1, no. 3, 2007, pp. 1–22.

[47] K. Kilkki, "Quality of experience in communications ecosystem," *J. Univ. Comput. Sci*., vol. 14, no. 5, pp. 615–624, 2008.

[48] M. Grega, L. Janowski, M. Leszczuk, P. Romaniak, and Z. Papir, "Quality of experience evaluation for multimedia services," Przegląd Telekomunikacyjny i Wiadomości Telekomunikacyjne, vol. 81, no. 4, pp. 142–153, Apr. 2008.

[49] NTT Network Technology Laboratories. (2014, Sept. 30). Communications Service QoE Assessment, Research Portal Site @ NTT [Online]. Available: http://www.ntt.co.jp/qos/qoe/eng/concept/directionality.html

[50] C.-C. Wu, K.-T. Chen, Y.-C. Chang, and C.-L. Lei, "Crowdsourcing multimedia QoE evaluation: A trusted framework," *IEEE Trans. Multimedia*. vol. 15, no. 5, pp. 1121–1137, Aug. 2013.

[51] VQEG eLetter. [Online]. vol. 1, no. 2. Available: http://www.vqeg.org