EVALUATATION OF THE EFFICACY OF SSIMWAVE SOLUTION TO AUTOMATICALLY ASSESS THE PRESERVATION OF CREATIVE INTENT OF VIDEO CONTENT DELIVERED TO CONSUMERS
[Draft Test Procedures – Work In Progress]

The American Society of Cinematographers
Motion Imaging Technology Council
and
SSIMWave Inc.

March 2016

This document describes the specifications of the equipment, materials, environment and procedures of the subjective testing to evaluate the preservation of creative intent of compressed video content delivered to consumers, as well as the efficacy of SSIMWave's solution to automatically predict subjective evaluations.[1]

1   Testing Materials, Equipment and Environment

1.1   Testing materials

The ASC will be responsible for providing and vetting all source material with image quality "as graded" in state-of-the-art color grading facilities and representative of the creative intent of the cinematographers and colorists involved. Pristine-quality test video materials will be provided that cover a wide range of content types, possibly including indoor and outdoor daylight and nighttime scenes, static and dynamic scenes, man-made architectures, humans, animals, plants, landscapes, and oceans, etc. Scenes will be included with controlled lighting where human skin tone reproduction can be evaluated. Scenes will be included to illustrate reproduction of smooth gradients across saturated and unsaturated colors as well as high levels of detail. Computer animations may be included if needed to illustrate specific aspects of HDR/WCG imagery.

The "As Graded" source materials will be 4K (4096x2160) or UHD (3840x2160) resolution and High Dynamic Range ("HDR")/Wide Color Gamut ("WCG") content in at least the digital cinema DCI P3 colour space or REC2020 colour space and at least digital cinema standard dynamic range or above. Scenes will be included containing colors outside of the P3 color space. Carefully calibrated scenes will be included that contain wide dynamic range, with specular highlights, levels of highlight detail, and levels of shadow detail that reflect the dynamic range capabilities of current digital cinema cameras. Files will be either in ACES or DCI standard formats, in RGB or XYZ color coordinates respectively. Frame rates will be limited to 60 frames per second or less, progressive scan only. Copyright issues should be cleared and permissions should be obtained for

---

[1]   For an overview of the ASC Technology Committee's SSIMWave evaluation, see the separate **Evaluation of the Efficacy of SSIMWAVE Solution to Automatically Assess The Preservation of Creative Intent of Video Content Delivered to Consumers** workgroup proposal document.

free use of the materials by the Working Group, for testing, reporting, and/or public demonstration purposes.

Short clips will be selected from the source material by SSIMWave and the ASC Technology Committee Working Group that are representative of narrative cinema and television content, and that can illustrate potential issues introduced by encoding and compression of that material during distribution and delivery to consumers. These clips will be used in the evaluation process described below.
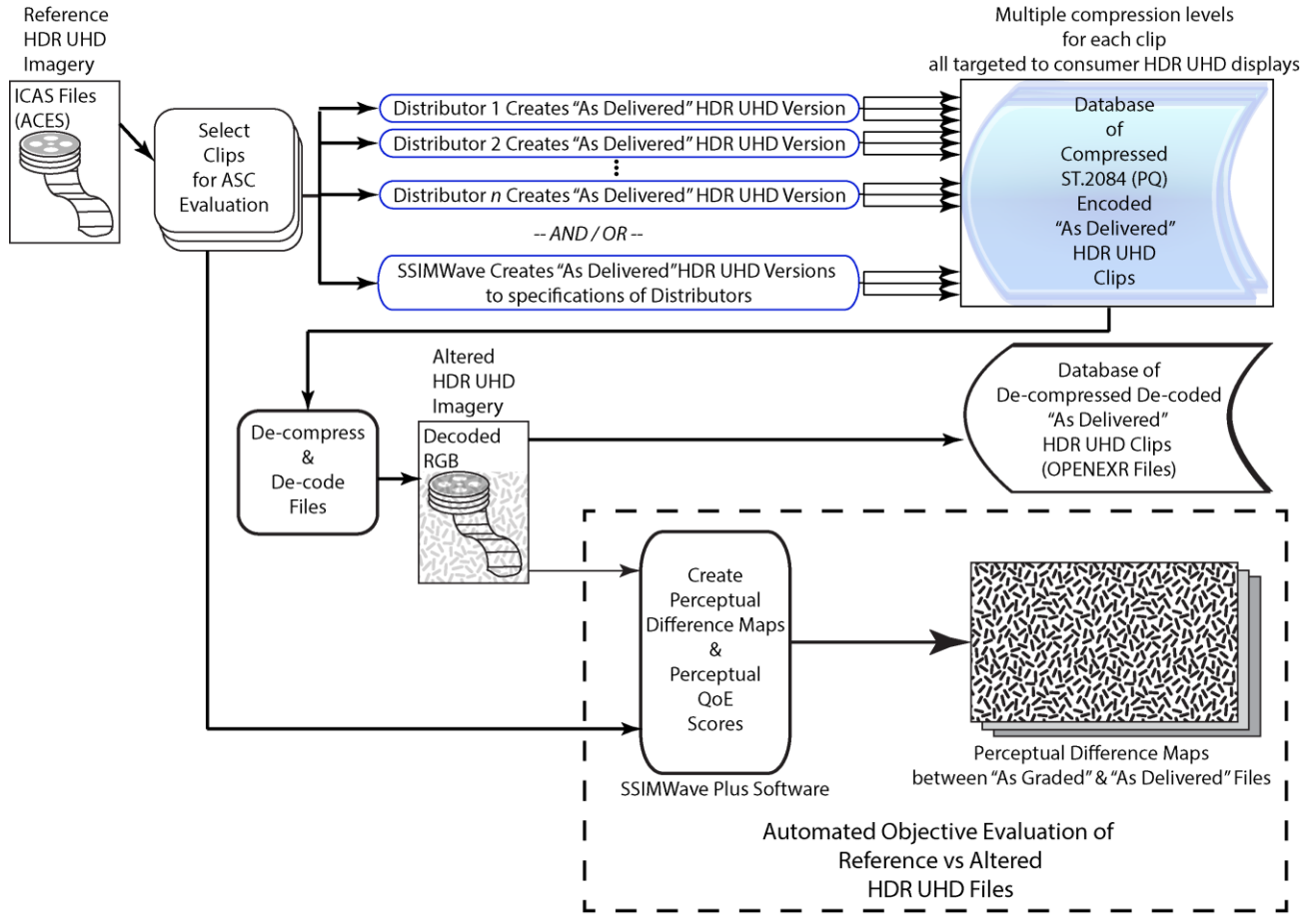
Test File Preparation & Objective Evaluation Stages

1.2 Test Material Preparation

A database of encoded, compressed versions of the test clips from the original source material will be prepared in advance for use in the evaluation.

There will be 20 source "As Graded" clips of approximately 10 seconds long each. The clips need to be selected to reflect the diversity of video content types. The source video clips will undergo the

stages of transcoding and compression corresponding to those performed during distribution and delivery of 4K/UHD, HDR/WCG material to consumer display devices.

Partner content distributors will be asked to specify the ingestion, encoding and compression steps used in their delivery of HDR, WCG, 4K or UHD content to their customers. "As Delivered" files will either be encoded and compressed by SSIMWave to the specifications of HDR UHD content distributors, or those distributors may choose to perform the encoding/encryption themselves and supply the "As Delivered" versions to SSIMWave.

In some cases, the original color-graded content may be encoded into "mezzanine" format video files for delivery from post house or color grading to distributors; such intermediate formats may include HDR UHD versions of ProRes or DNxHR. These "mezzanine" files will then be further transcoded and compressed according to that distributor's real-world HDR/WCG distribution workflows. In other cases, distributors may accept (or require) RAW or ACES files directly from content producers, and ingest these files directly into their workflow. Original, source test material will likewise be processed, transcoded and/or compressed to reproduce any video degradation that occurs during that transfer and ingestion process of a given distributor.

The compression configuration details from the industrial video distribution partners will be adopted for the specific setup of the workflow. If content producers are unwilling to divulge proprietary information to SSIMWave even under NDA, or if the software necessary to reproduce those steps is not available to SSIMWave, then content producers will be asked to perform such encoding and compression on the source test material and return the resulting "As Delivered" files for use in this evaluation.

The encoders may follow MPEG-2, H.264 or HEVC standard, with ST.2084/Perceptual Quantizer (PQ) electro-optical transfer function.[2] Each "As Graded" reference clip will be encoded with at least four (4) levels of compression that span the bit-rates used by distributors when targeting consumer HDR, UHD displays. "As delivered" files may contain content targeted to either DCI P3 or REC2020 color space, depending upon the conventions and standards of the distributor, regardless of the color space of the original source material.

It needs to be mentioned that in practice, distributors may create many versions of "As Delivered" files as derivatives from the "As Graded" reference file, many of which may be down-scaled versions that have lower resolutions spatially and/or temporally. Our interest in the current test phase is in those "As Delivered" files that maintain the same spatial resolution (i.e., 4K or UHD) as the reference videos.

Likewise, only levels of compression representative of those typically used to deliver HDR, UHD/4K content to consumer HDR UHD displays will be evaluated. Encoding and compression targeted to other (e.g., "mobile") types of consumer devices will not be used during this initial evaluation.

---

[2] Hybrid Log Gamma (HLG) HDR encoding will not be evaluated at this time. The initial ASC evaluation will use only ST.2084/PQ encoded "As Delivered" material. HLG encoding may be evaluated in a subsequent phase specifically to compare differences between these two transfer functions.

The subjective evaluations will compare original "as graded" source material against the same material "as delivered" to consumers, to evaluate the extent to which the original creative intent is preserved or altered, and the extent to which SSIMWave quality metric software reflects those perceived alterations. The level of degradation caused by individual intermediate steps – while important – will not be evaluated by cinematographers during this initial evaluation.

Assuming that HDR content distributors target either DCI P3 or REC2020 color space for their delivered content and not both, and also assuming two independent distributors supply their "As Delivered" videos together with SSIMWave versions of "As Delivered" videos, a total of 240 compressed "As Delivered" video files will be available for comparisons with their corresponding "As Graded" versions.

All "As Delivered" video clips will be fully decompressed and stored as OpenEXR format files (in ACES color space, and with a standard ST.2065.4 container but with AcesImageContainerFlag set to 0) prior to evaluation. This allows the image data to be stored as RGB floating point numbers that can be played back on any monitor with an ODT. This also avoids the quantization issues in any integer-based color encoding approach that may alter the true color of the pixels created by the decompression process.

The decoding of "As Delivered" clips will take into account metadata settings that a given industry partner sends to consumer monitors when delivering their HDR WCG content to such displays.

1.3   Objective Testing

SSIMWave will assess all compressed videos against the original source material using its SQM software. The results should not only include per-frame and per-video quality scores, but also pixel-precision quality maps that indicate local quality variations across space and time. The quality maps will be computed and stored as video files. To have better control of playback smoothness in the subjective testing, SQM software will not be used to compute quality scores and quality maps on site. Instead, the quality map videos precomputed by SQM software and stored as video files will be used in the subjective testing for expert verification.

1.4   Testing environment and equipment

Environment: laboratory environment with low room illumination.

Displays:
- Two identical HDR/WCG, 4K or UHD resolution reference monitors (for side by side comparison of source "As Graded" vs. compressed "As Delivered" videos),
- One UHD monitor (to show the SSIMWave quality map, as a prediction of local video quality). This monitor will be a "typical" high-end HDR UHD consumer display (also used for illustrating to participants the likely differences between reference-grade displays and actual customer display devices).
NOTE: The consumer display will not be used for evaluation of delivered quality.
All monitors will be capable of displaying at least UHD resolution at the full frame rate of the original source material.

Computer:
- PC with high-performance graphics card, and with sufficient CPU power, RAM and hard drive space and transfer rates, and output interface to drive the playback of the source and test videos on the reference monitors.
- PC capable of displaying the SSIMWave SSIMPlus software's QoE metric and visual quality disparity map on a separate monitor simultaneously with the display of the source and test videos.

Software: able to play the HDR/WCG UHD/4K video on the reference monitors, and synchronize the source and test videos and quality maps.

2   Subjective Testing Process

Subjective Evaluation Stage
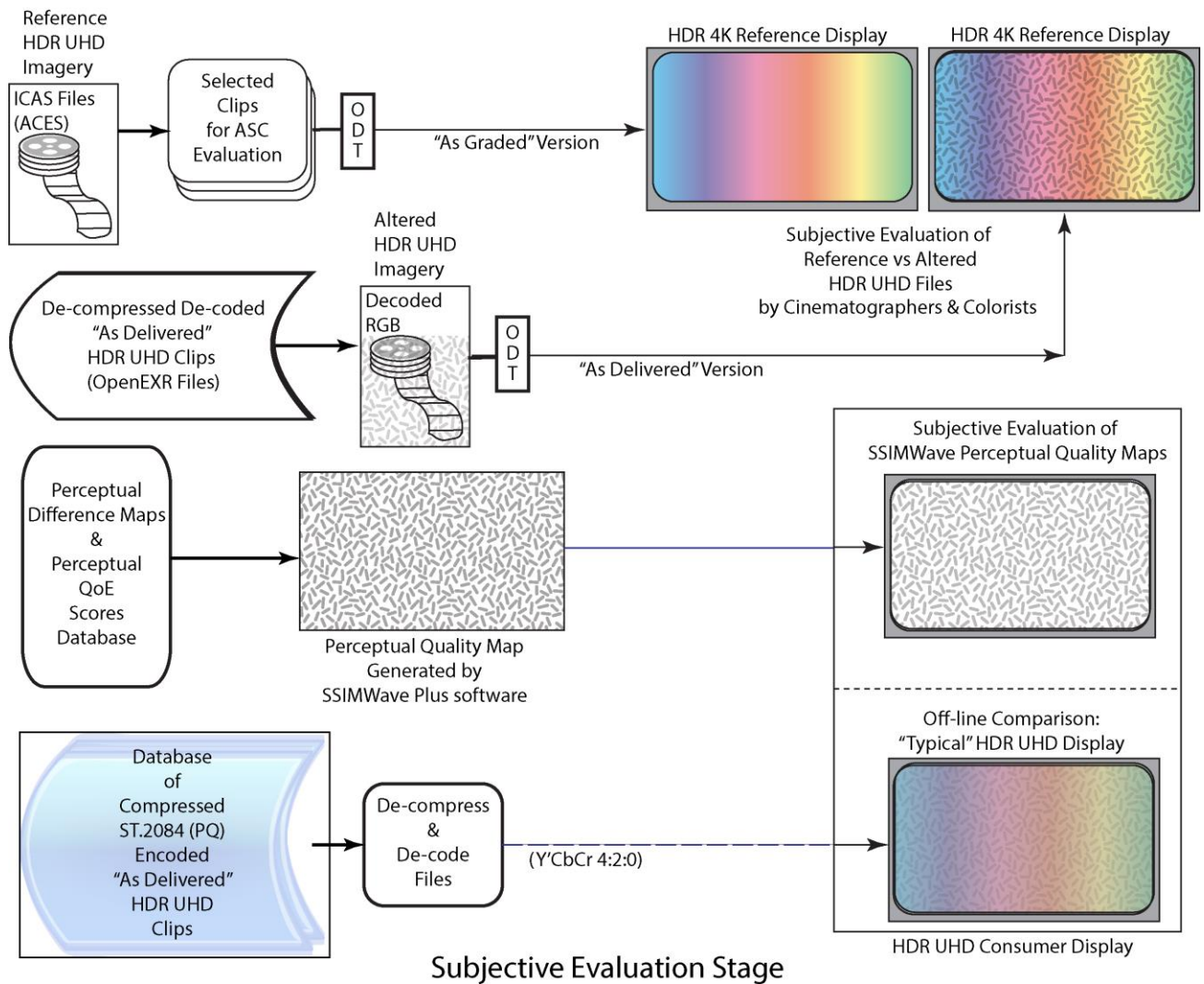
2.1   Subjects

A total of at least 25 to 30 cinematographers, colorists, or industrial "golden eyes" will be recruited to participate in the subjective testing.[3] Background information such as occupation, gender, and age will be documented.

2.2   Testing method

1. Two reference monitors are used to show the source ("As Graded") and test (decompressed/decoded "As Delivered") videos side by side. The display of the two monitors are synchronized. The subject(s) rate the degree to which the original creative intent of the "As Graded" content is preserved by the "As Delivered" content. Viewers are allowed to play the videos, and also have the choice to pause the playback, and stop at any time instance to have a close observation and comparison of the local details of the source and test video content. There is no constraint on the viewing distance, so that the subjects can get as close to the screens as possible.

2. The same "As Graded" and "As Delivered" video clips are re-played on the two reference monitors while a third monitor shows the corresponding quality map video generated by SSIMWave's QoE measurement tool. The display of the 3 monitors are synchronized. The same control of video playback is available to the subjects. The subjects rate the degree to which the SSIMWave quality map reflects their subjective judgment of the preservation or degradation of the "As Delivered" content vs. the original "As Delivered" content.

3. After the evaluation of a clip is completed, the subjects may elect to view the "As Delivered" clip on the "typical" consumer UHD HDR monitor, as a sample of how such a consumer display may alter the content vs a reference monitor. This is for the subjects' information only. This test is not designed to evaluate or rate consumer monitors, which vary from brand to brand, manufacturer to manufacturer, model to model, and from one user-settable menu setting to another. The actual monitor used may vary depending upon availability.

---

[3] NB.: At least 30 participants are required to ensure adequate sample size for statistical validity of the evaluation, as per normal testing practice.

Fig. 1 Layout of answer sheet. "A" is the probability of release, "B" is the meaningful measure of SSIMWave's quality map.

A layout of the answer sheet can be seen in Fig. 1. The subjects are required to give at least two responses using two sliders. Each slider is divided into 5 segments marked from right to left (or top to down) by "Excellent", "Good", "Fair", "Poor", and "Bad", respectively. The slider can record a score between 0 and 100, equally divided according to the 5 segments above.

The first response of the subjects is on "How well does 'As Delivered' content match the original 'As Graded' content, in terms of your likelihood of approving the "as delivered" version for release? Would you approve the test video to be released as preserving the creative intent of the source video?". The definition of the answers for each segment can be found below.

- "Excellent": almost certain to release: very close match to original, suitable for release to even high-quality HDR/WCG display devices and viewing environments;
- "Good": would probably release: a few noticeable degradations from the original, but sufficiently close for release to consumers on normal to high end HDR/WCG consumer displays;
- "Fair": marginal probability of release, lots of noticeable degradations such that original creative intent is compromised and the image is not suitable for general release to HDR/WCG consumer displays;
- "Poor": low probability of release, degradations severe enough that original creative intent is highly compromised and the image is not suitable for release to HDR/WCG consumer displays;
- "Bad": no probability of release: degradation so severe that artifacts draw attention to themselves. Should never be released to consumers.

The second response is on "How well does SSIMWave's quality map predict the level of the preservation of creative intent?" The definition of the answers for each segment can be found below.

- "Excellent": the SSIMWave's quality map makes perfect sense in terms of the level of the preservation of creative intent;
- "Good": the SSIMWave's quality map is able to mostly predict the level of the preservation of creative intent;
- "Fair": the SSIMWave's quality map is matching around half of subject's opinion in terms of predicting the level of the preservation of creative intent;
- "Poor": the SSIMWave's quality map is able to predict the level of the preservation of creative intent at few locations in a video;
- "Bad": the SSIMWave's quality map doesn't make sense at all.

In addition to the required responses, the subjects would have options to tick specific predefined issues listed in Appendix A, as well as the ability to add as many additional concerns and their importance as desired in a comment window.

2.3   The test procedure

Before the formal subjective testing, the subjects are given general instructions on the purpose of the test and the procedures, on how to use the sliders to give scores, and on how to input comments. They are also informed of the questions to be asked during the test. Sufficient time is allowed for visual adaptation of participants to the darkened viewing environment of the testbed setting. Eight (8) sample "training" videos will be used at the beginning of the test to help the subjects get familiar with the testing procedure.

During the formal test, the test videos will be displayed in a random order. Ten percent (10%) of the videos will be repeated. Such data can be used to evaluate the subject's consistency in evaluating the video quality. When evaluating a test video, the subject will have the control to play the videos, or to stop at any frame, and compare the corresponding frames of the source and test videos, and compare the corresponding quality map given by the SSIMWave QoE measurement tool.

At most three participants can view the content simultaneously, with one responsible for controlling the playback. The participants must be seated such that angle of view and viewing distance to the reference monitors and SSIMWave software monitor are not compromised. If more than one participant is evaluating material simultaneously, they must agree not to share comments or interpretations, and they must be monitored to ensure that their evaluations remain independent.

In order to avoid fatigue effect of the subjects, the total test for each subject may be divided into multiple sessions. A single session should last no more than 20-25 minutes. Several sessions may be done in a row, but there should be at least a five (5)-minute break in between.

Each participant should expect to spend no more than two (2) hours performing the evaluation.

Appendix A   Specific Video Artifacts that May Affect the Preservation of Creative Intent

The purpose of collecting feedback on the specific types of video artifacts in the subjective testing is to have a better understanding about how cinematographers, colorists, or industrial "golden eyes" make judgments on the preservation of creative intent, and also to identify the major quality issues and assess the relative severeness/importance between them in practical encoding processes during video distribution. Therefore, the following questions will be asked in the subjective test in addition to the questions of "How well does "as delivered" content match the original "as graded" content", and "How well does SSIMWave's quality map predict the level of the preservation of creative intent".

NOTE: *It is recognized that assigning a single arbitrary number or quality rating to an image is not something that is natural to professional colorists and cinematographers. The following is an initial attempt to indicate issues noticeable during the evaluation in ways and using terminology common in the industry.*

However, one needs to be aware that the lossy encoding process in video distributions may create many different types of artifacts. In addition to the criteria listed below, other types of artifacts appear – some of which may not be readily classifiable in the terms commonly used by industry professionals. On the other hand, some of the criteria in the list may not be a major quality issue in the video distribution process. Therefore, the following list will need to be adjusted or redefined – including by showing sample "As Delivered" videos to expert viewers and asking for their feedbacks prior to the formal test phase.

How significantly did alterations or artifacts to any of the following affect your decision, positively or negatively?

1. Highlight detail
2. Shadow detail
3. Noise
4. Exposure shifts
5. Color cast/consistency
6. Skin tone
7. Banding/Moiré
8. Additional comments (define your own criteria and describe your observation)

NOTE: *The above is subject to change. Comments and suggestions are encouraged.*

Appendix B   Protection of Proprietary Information and Intellectual Property

1. Test Material

Test imagery used in this evaluation will be used for the internal purposes of the ASC only. All display and distribution of such material to industry partners is to be done solely for the purposes of preparing for and/or carrying out this evaluation under the auspices of ASC Technology Committee. Otherwise, there are no other restrictions to those allowed to view test content nor to allowed modifications of the content for the purposes of this study.

2. Proprietary Information

Proprietary information and intellectual property of industry partners will be kept confidential by SSIMWave Inc. during the preparation for and execution of this study. Non-disclosure and limited use agreements can be negotiated between SSIMWave Inc. and cooperating industry partners where necessary. Test subjects, members and associates of the ASC, and other volunteers who help in the planning, preparation, and execution of the evaluation will not be required to sign NDA or other use restriction agreements as a result of participating in this study. Care will be taken that participants in the study, members and associates of the ASC, and volunteers are not exposed to information proprietary to any organizations cooperating in this project.

3. No comparisons between distributors of HDR, Wide Color Gamut, and UHD Content

The goal of this study is to evaluate whether the SSIMWave software accurately reflects what cinematographers and colorists notice in terms of preserving HDR, wide-color, 4K/UHD image quality, and how useful it could be in the production and post-production process, not in ranking content providers or distributors. It is not to evaluate the "quality" of one particular distributor's delivered HDR, UHD content vs that of another distributor. The ASC will not endorse, promote, or recommend any one distributor over another as a result of this study, and no report or presentation that results from this study will do so.

No information as to which protocols and processes are used by a given distributor will be known outside of possibly SSIMWave Inc. and the industry partner themselves. Only SSIMWave, under NDA, would ever know which test files came from a given distributor, or were processed according to a particular distributor's instructions. No ASC members, personnel, or associates, and no participants in the study, will know -- or be able to infer -- source(s) of the encoding or compression mechanisms used on any given shot or clip used in this test, either before, during, or after the evaluation by cinematographers and colorists. SSIMWave has agreed to ensure that such information will not be disclosed, either to the ASC or publicly, and will take the necessary steps to ensure that such information cannot be indirectly inferred from published results.

It is possible that professional journal reports or industry presentations based on this study could include differentiating between different technologies used in the post-processing and distribution pipeline, such the use of HEVC, etc. The ASC will insist, and SSIMWave has agreed, that any such study or report be done only in such a way that the use of any specific technologies by any given

distributor will be obfuscated -- e.g., by using for such purposes open source or commonly available versions and commonly accepted configurations of such technologies, in addition to those used by any specific distributor or network. Likewise, all source content used as a "reference" in quality comparisons will be from or vetted through the ASC (such as ICAS or StEM files); no test material will be traceable to any particular industry partner.

If an industry partner wishes to privately obtain results specific to them, vs. the overall, aggregate industry results that will be made public, as a benefit of their participation in this evaluation, that can be arranged directly between that partner and SSIMWave Inc. No public reports nor reports to the ASC would contain such partner-specific information.