INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO-IEC/JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO

ISO-IEC/JTC1/SC29/WG11
N2823
July 1999

Source: **Test and Video Groups**
Status: **Draft**
Title: **Report Of The Formal Verification Tests On MPEG-4 Temporal Scalability in Core Profile**

Summary

This report illustrates the results of the verification test to evaluate the performance of MPEG-4 video Temporal Scalability tool in the Core Profile.
The test was performed using the "Single Stimulus" method. This method particularly suited to evaluate the features of multimedia video signals. The subjects were required to evaluate the annoyance of impairments of compressed sequences with and without temporal scalability functionality. The test was conducted using a total of 45 subjects in two different laboratories and the results showed that the quality of sequences encoded using MPEG-4 temporal scalability tools are comparable to the quality of sequences encoded without temporal scalability.

1. Introduction

The visual part of the MPEG-4 standard will provide a toolbox containing tools and algorithms bringing solutions to a number of functionality and covering a wide range of bitrate.
It was recognised that the verification tests should address functionality and applications that are potentially of great interest for users. The scalability has been identified as one of main functionality and temporal scalability in core profile is successively verified after the temporal scalability test in simple scalable profile.
It is the purpose of this document to describe test procedures and outcomes of the temporal scalability test in core profile.
The test was carried out in two laboratories: FUB and CSELT.

2. Context and test motivation

2.1. Temporal scalability in core profile

Temporal scalability bitstreams consists of multi-layer bitstreams, for example, a base layer bitstream and an enhancement layer bitstream. The base layer bitstream provides basic frame rate. The enhancement layer bitstream provides the reconstructed frames in different display time from the base layer. By decoding both the base and enhancement layer bitstreams, reconstructed frames with higher frame rate are provided. Moreover, the temporal scalability in core profile bitstreams can contain an arbitrary shaped video object. For this test, both the base layer and the enhancement layer comply with the core profile. These profiles were selected such that core profile decoders can decode the base and enhancement layer bitstreams.

2.2. Anchors

Single layer and simulcast coding in core profile are used as two anchors. The single layer coding uses the same bitrate as the total bitrate of two bitstreams with temporal scalability. The frame rate of single layer coding is as same as the base layer + enhancement layer. The single layer coding provides a single bitstream, and therefore it has less functionality than temporal scalability.
The simulcast coding generates two bitstreams. One bitstream has lower frame rate. And the other has higher frame rate. The lower frame rate bitstream has a same frame rate as the base layer of the temporal scalability. The higher frame rate bitstream has the same frame rate as the base layer + enhancement layer. The simulcast coding provides same functionality as temporal scalability does and the picture quality of lower frame rate bitstream is almost as same as the base layer of temporal scalability bitstreams. Therefore the higher frame rate results from simulcast coding to be compared to the temporal scalability.

2.3. Test motivation

The tests have been designed to verify the following two facts.

- The quality provided using temporal scalability in core profile is similar to the quality provided by single layer coding in core profile.
- The quality provided using temporal scalability in core profile is better than the quality provided by the simulcast coding in core profile.

3. Time Schedule

The actual schedule of activities involved in the subjective video temporal scalability test in core profile, as well as the organisation conducting each phase of the work is listed in the table below. This was finalised in the Vancouver meeting. The test materials were encoded by Sharp and sent to FUB. The tapes were edited by FUB. The subjective test was conducted by FUB and CSELT.

Tasks	*Time*	*Output*	*Responsible*
Generate D1 tapes and distribute them to FUB	5 weeks	April 29th	Sharp
First draft of the report	3 weeks	May 17th	Sharp
Test tape edit	-	May 17th	FUB
Formal subjective tests	2 weeks	May 31st	FUB CSELT
Report	2 weeks	June 14th	FUB Sharp

Table 1: Schedule for the formal test of the TSCP

4. Test Conditions

In Seoul, the pre-screening shown satisfying results. It was decided to define the test conditions as depicted in the table here below:

Test Codec	MPEG-4 CP, temporal scalability arbitrary-shaped P-VOP
Anchors	1. Object based single layer: MPEG-4 CP one layer, same bit rate and frame rate as base + enhancement layer. 2. Object based simulcast: MPEG4 CP one layer, same bit rate as enhancement layer and same frame rate as base + enhancement layer.
Test Sequences	Sean (object: man) News (object: monitor) Bream (object: fish) Singer (object: singer) Dancer (object: dancer)
Bitstream Encode	Sharp
Coding Parameters	Refer to Table 3
Other Coding Parameters	• OBMC : disable • change CR : disable • quant type : H.263 • Error resilence : disable • DC/AC pred : enable • Quantization : Fixed Qp • alpha threshold: 0 (lossless coding) • input frames : 0 – 299 • enhancement type : 0 • VOP_coding_type : I,P and B VOPs are used in Anchor tests (single layer and simulcast). : I and P VOPs are used in CP test. • Deblocking filter : enable

Table 2: Temporal scalability in core profile test conditions

Note: coding type "I,P and B-VOP" implies encoding the 1st frame as I-VOP and the others as P and B-VOP.

4.1. Test Sequences

Five test sequences, "Sean", "News", "Bream", "Singer" and "Dancer" are used for the verification test. Each sequence has an associated alpha mask and this was used to distinguish foreground object and background object. "Sean" has relatively static motion, and is low complexity sequence. Other sequences have a large motion, and are high complexity sequence.

4.2. B-VOP

B-VOP coding tool is included in core profile but NOT included in temporal scalability in core profile. Therefore B-VOP coding tool is available only in single layer and simulcast coding, and is not used in temporal scalability coding.

4.3. Rate Control

The fix Quantising Parameter value is used. In order to match the target bitrate, appropriate the Quantising Parameter value and frame rate are chosen. Target bitrate are defined among pre-defined four kinds of bitrates which come from a bandwidth of real communication environment.

4.4. Coding Parameters

The formal test has been conducted according to the conditions described in document N2666, and also reported in the table here below.

Sequence	condition	Base layer	Enhancement layer	Single layer (Anchor1)	Simulcast (Anchor2)
News	frame rate	2.5	7.5	7.5	7.5
	target bit	12	12	24	12
	resolution	QCIF	QCIF	QCIF	QCIF
	frame rate	3.33	10	10	10
	target bit	24	24	48	24
	resolution	QCIF	QCIF	QCIF	QCIF
	frame rate	5	15	15	15
	target bit	48	48	96	48
	resolution	CIF	CIF	CIF	CIF
Sean	frame rate	2.5	7.5	7.5	7.5
	target bit	12	12	24	12
	resolution	QCIF	QCIF	QCIF	QCIF
	frame rate	3.33	10	10	10
	target bit	24	24	48	24
	resolution	QCIF	QCIF	QCIF	QCIF
	frame rate	5	15	15	15
	target bit	48	48	96	48
	resolution	CIF	CIF	CIF	CIF
Bream	frame rate	2.5	7.5	7.5	7.5
	target bit	12	12	24	12
	resolution	QCIF	QCIF	QCIF	QCIF
	frame rate	3.33	10	10	10
	target bit	24	24	48	24
	resolution	QCIF	QCIF	QCIF	QCIF
	frame rate	2.5	7.5	7.5	7.5
	target bit	48	48	96	48
	resolution	CIF	CIF	CIF	CIF
Singer	frame rate	2.5	7.5	7.5	7.5
	target bit	12	12	24	12
	resolution	QCIF	QCIF	QCIF	QCIF
	frame rate	3.33	10	10	10
	target bit	24	24	48	24
	resolution	QCIF	QCIF	QCIF	QCIF
	frame rate	5	15	15	15
	target bit	48	48	96	48
	resolution	CIF	CIF	CIF	CIF
Dancer	frame rate	2.5	7.5	7.5	7.5
	target bit	24	24	48	24
	resolution	QCIF	QCIF	QCIF	QCIF
	frame rate	3.33	10	10	10
	target bit	32	32	64	32
	resolution	QCIF	QCIF	QCIF	QCIF
	frame rate	2.5	7.5	7.5	7.5
	target bit	48	48	96	48
	resolution	CIF	CIF	CIF	CIF

Table 3 Coding parameters for verification test

In any test, different conditions were applied changing the frame rate, the bit rate and the resolution. In the table, frame rate of the enhancement layer indicates the frame rate of base + enhancement layer. The foreground object and background object is encoded as separate VO. They are composed by piling up these objects. Only a foreground object is encoded by temporal scalability coding or anchor coding and the background object is commonly used.

Different test sequences were used according to their coding difficulty and with the aim to cover the wider number of representative cases. This experiment was performed under the assumption that MPEG4 temporal scalability will be used for real applications. Therefore, the test condition places a restriction on bitrate alone. Other coding parameters, frame rate, resolution, and coding type are set by sequence characteristics. In real situations where MPEG-4 temporal scalability will be used, contents providers will select their target bitrate from existing network capacities (e.g. modems with analogue telephone lines, ISDN, CATV, etc). Under similar test conditions, temporal scalability in core profile produced good results. For this reason, the test conditions shown above have been applied for the formal test temporal scalability in core profile.

4.5. Display format

The test material generated by the coding processes, and the original test sequences, will be up-sampled to the 601 format at 60Hz (720x486).
The actual size of the up-sampled CIF sequences is 704x576; to fit this format into a 720x486 frame the following operation will be made:

- 8 bit wide black pixels will be added before and after each 704 line;
- The first and the last 45 lines of the 704x576 frame will be deleted.

5. The formal subjective test

The formal subjective tests have been conducted in CSELT and FUB using the Single Stimulus test method with a total of more than 45 subjects.
The SS (Single Stimulus) test method test method has been selected as the more adequate to conduct the test of the temporal scalability.

5.1. Single Stimulus (SS)

Experts in subjective quality assessment have seen the test materials in Dublin, Atlantic City, Rome and Seoul in order to be able to propose a suitable methodology for the subjective tests. The assessment of the encoded sequences on the point of view of the overall perceptual image quality led to the proposal of the Single Stimulus (SS) method. Indeed, since the aim of the test is to compare the quality of the encoded/decoded sequences using different algorithms, it is recommended to make the subjective evaluation using a quality scale.
At the condition under consideration, it has been noticed that the encoded/decoded sequences are very impaired with respect to the source in CIF format. Therefore, since a comparison method does not provide stable results when the two sequences to compare are too differently impaired, the use of a simple stimulus is advisable instead.
In the SS method an 11-point quality scale is used, ranging from 0 to 10 in steps of size 1, with the best score being 10. Test subjects vote once, giving their assessment of the quality of the test sequence presented.
The instructions to be given to the observers are detailed in appendix A.

5.2. Laboratory set-up

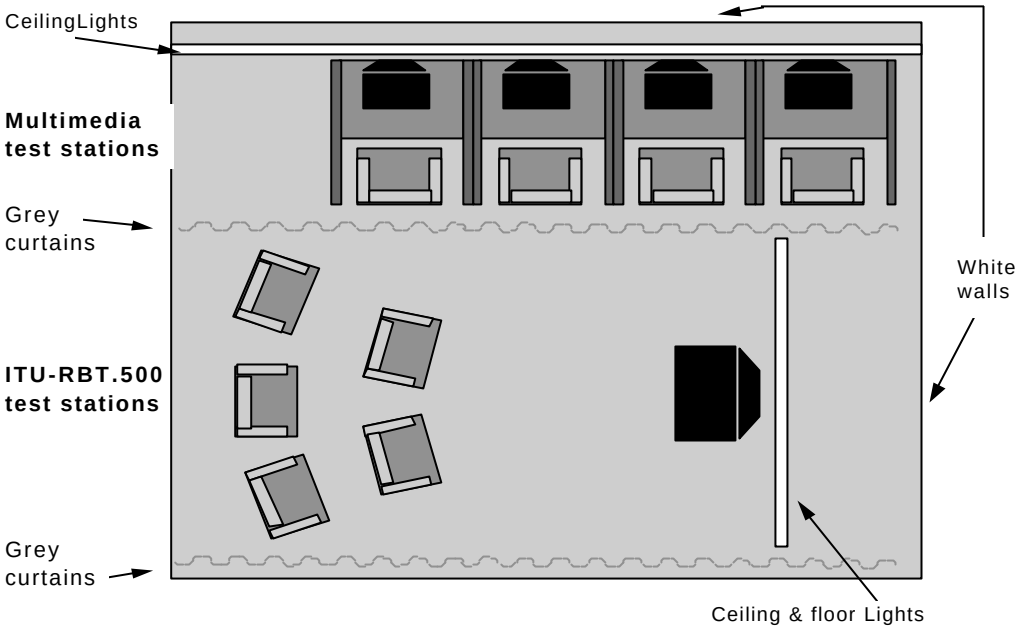
The set-up of the test laboratories has to be done according to the recommendations BT.500-7 (ITU-R). One or more 19" professional monitor was used (e.g. Sony BVM1910) to display the sequences.
The viewers were located in front of the monitor at distance equal to 4 times the height of the screen.
Each test was done using a panel of 9 non-expert viewers. Three of them were located in front of each monitor.

5.2.1. Laboratory set-up at CSELT

For this test two rooms were used: the 'control' room and the 'test' rooms.

- The 'control' room hosted the VTRs and the PC that controls both VTR playing and data collection.
- The 'test' rooms host test stations with suitable signal transducers (e.g. monitors, headphones, loudspeakers, telephones, etc.)

During the test, subjects performed their task in the 'test' rooms, while an operator was monitoring, from the 'control' room, their behaviour through a closed circuit television system and check the correct running of the test.
The layout of this room is shown on the following figure. Ceiling, walls and floor are covered by sound proofing material, in order to keep the reverberation under the thresholds specified by ITU-T Rec. P800.



5.2.1.1. The equipment

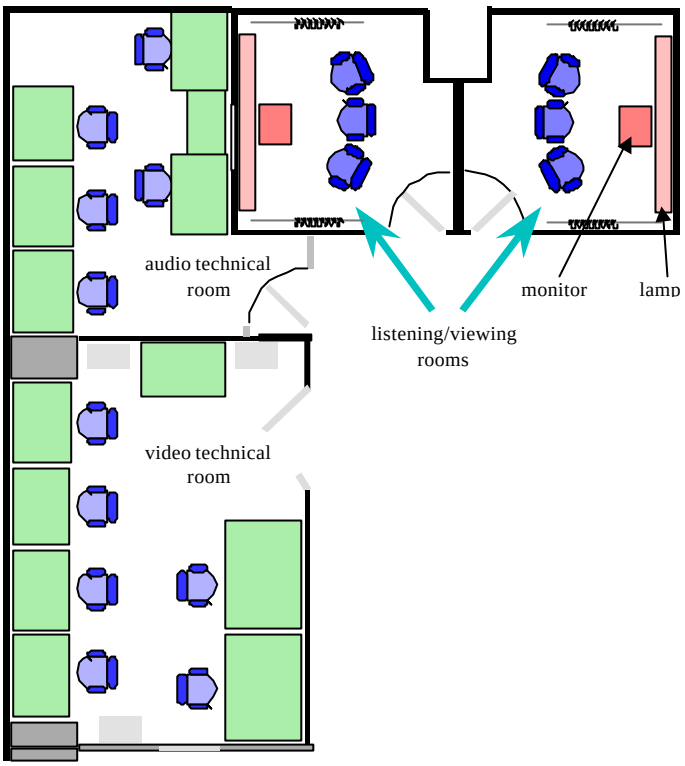
Test conditions were recorded on a Rewritable Laser Disk (Pioneer VDR-V1000). This is the system that is commonly used for the test, because with this system test conditions can be presented in several different orders, by controlling the recorder from a PC through a RS232 interface. This feature was not used in this test, to avoid source of variations between results of different laboratories.
The sequences were displayed on a Sony BVM2011monitor (20 inches), that was previously calibrated according to [7].
Votes were collected automatically by using three Penny & Giles PGF5000 5 KOhm sliders. Subjects are pre-screened for visual acuity and colour blindness by using a Monoyer Optometric Table (3 m. viewing distance) and Ishihara's tables respectively.

5.2.1.2. The software

The software for the automatic collection of votes and for the control of the Laser Disk has been developed in CSELT by a sub-contracted expert in software development.
The data are organised both by using the format specified by the ITU-R and under a format particularly useful for building a data base of test results.
The same expert is now developing modules for experimental design and statistical analysis. Both these functions are based on procedures of SAS, a powerful package of statistical analysis, that, besides the commonly used experimental designs (completely randomised and block randomised) and statistical analysis, will allow more sophisticated and powerful investigations.

5.2.2. Laboratory set-up at FUB

The subjective assessment environment consists of four different rooms. The first and the second rooms are the «listening/viewing room» where tests are performed, the third one is the «audio technical room» containing a digital audio recorder and the other equipment necessary to perform audio tests and the fourth one is the «video technical room» containing the DVTR two video disks and the other equipment necessary to perform tests video and audio visual tests.



FUB's audio/video subjective assessment laboratory

5.2.2.1. The equipment

Test sequences were recorded on four D1 tapes and displayed to the observers on two Sony BVM 20E1/E grade A professional studio monitor using a D1 BTS/Philips DVTR controlled by the IQ++ platform including hardware and software and described in the previous section. The monitors were calibrated using a Sony custom probe and the standard Pluge signal.
Votes were collected using the IQ++ sliders.
Six groups of 3 non-expert observers were participated to the test, organised into 2 sessions. Observers were different for 50 and 60 Hz tests. Prior to the test, subjects were screened for visual acuity by using a Monoyer Optometric Table. Besides, test for normal colour vision was performed using Ishihara's tables. These tables are designed to detect colour blindness or strong colour vision deficiencies; that is to say, to check ability of the observers in discriminating colours. The subject who is not able to pass the test is discarded.
All the viewing conditions were compliant with the ITU-R B.T. Rec. 500.

5.2.2.2. The software

The software platform is the IQ++, developed and commercialised by CCETT and described in previous section.
The results have been further processed using Microsoft Excel to obtain some comparison of the influences among session and presentations.

5.3. Experimental design of the test

The sequences were recorded on a D1 tape one after the other without any interruption between two subsequent sequences. One presentation order was prepared since the contextual effect in SS method has a minor impact on the reliability of the results.

6. Analysis of the results of the test

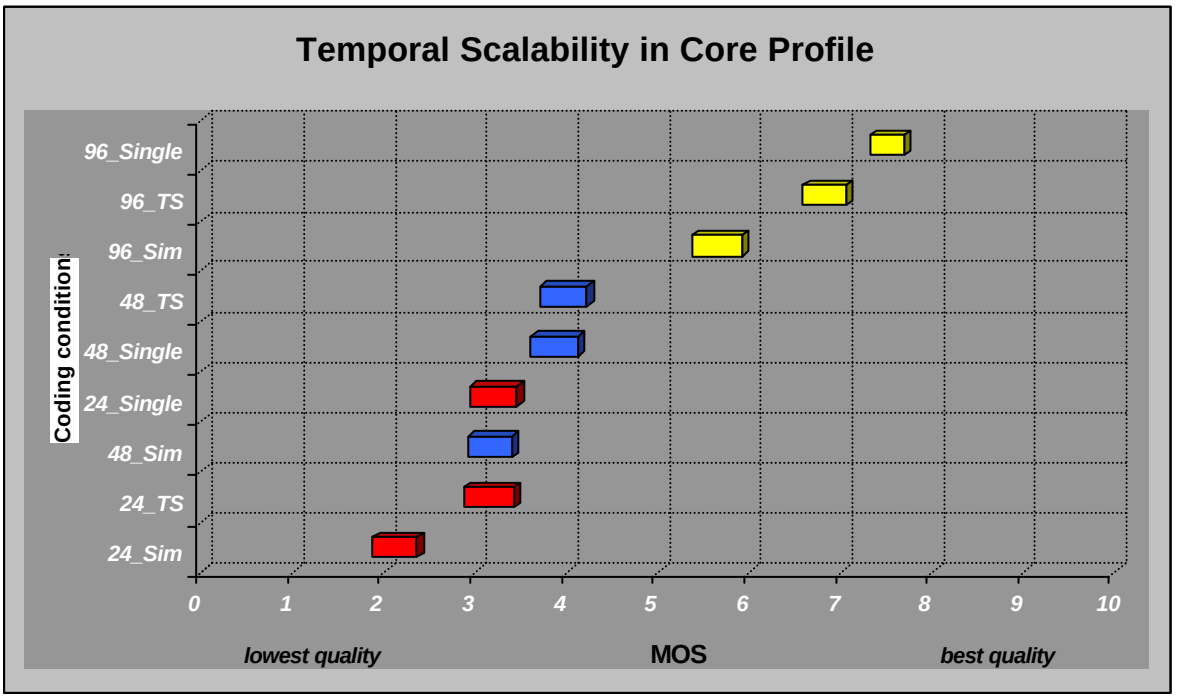
Table 4 provides the detailed test results for each sequences and coding conditions; Table 5 provides the overall results of the test..

	Dancer	*News*	*Sean*	*Singer*	*Bream*	*GENERAL*	
24_Sim	**1,33**	**1,33**	**1,11**	**1,17**	**4,44**	**1,88**	***Mean***
	1,20	0,76	0,77	0,64	1,10	0,48	*C.I.*
24_TS	**2,33**	**1,89**	**2,89**	**2,83**	**4,44**	**2,88**	***Mean***
	1,35	0,84	1,60	1,03	0,89	0,55	*C.I.*
48_Sim	**1,44**	**2,89**	**2,94**	**2,72**	**4,61**	**2,92**	***Mean***
	0,84	0,99	1,36	0,77	0,98	0,49	*C.I.*
24_Single	**2,44**	**1,78**	**2,44**	**3,00**	**5,06**	**2,94**	***Mean***
	0,81	0,89	0,97	1,15	1,28	0,51	*C.I.*
48_Single	**2,94**	**2,89**	**3,22**	**3,06**	**5,94**	**3,61**	***Mean***
	1,40	0,81	0,89	0,94	1,00	0,51	*C.I.*
48_TS	**2,22**	**3,94**	**3,72**	**3,72**	**4,94**	**3,71**	***Mean***
	1,09	1,44	0,80	0,96	1,10	0,51	*C.I.*
96_Sim	**2,17**	**6,00**	**4,44**	**6,78**	**7,50**	**5,38**	***Mean***
	0,87	0,87	0,87	0,62	1,02	0,54	*C.I.*
96_TS	**4,39**	**6,11**	**6,56**	**7,33**	**8,56**	**6,59**	***Mean***
	0,88	0,72	1,08	0,85	0,58	0,47	*C.I.*
96_Single	**5,28**	**7,11**	**7,83**	**7,78**	**8,61**	**7,32**	***Mean***
	0,54	0,76	0,71	0,58	0,64	0,37	*C.I.*
GENERAL	**2,73**	**3,77**	**3,91**	**4,27**	**6,01**	**4,14**	***Mean***
	0,39	0,43	0,46	0,45	0,41	0,20	*C.I.*

Table 4 – Detailed results

	Simulcast		***Temporal Scalability***		***Single layer***	
	Mean	C.I.	*Mean*	C.I.	*Mean*	C.I.
24 Kbps	***1,88***	0,48	***2,88***	0,55	***2,94***	0,51
48 Kbps	***2,92***	0,49	***3,71***	0,51	***3,61***	0,51
96 Kbps	***5,38***	0,54	***6,59***	0,47	***7,32***	0,37

Table 5 – Overall results



Graph 1 – Temporal Scalability in Core Profile – Test Results

7. Conclusions

The formal verification test showed that in two of the given conditions, the Temporal Scalability in Core Profile exhibits the same overall quality provided by Single layer coding in Core Profile; for the third case the quality value are very close.
Furthermore it is evident that the Temporal Scalability in Core Profile provides better quality than the Simulcast coding in Core Profile.

8. Acknowledgements

The authors of this report would like to thank the following additional people, and their companies, for their contributions to the completion of these tests.

Preliminary studies to produce pre-screening material	H.Katata, N.Ito	Sharp
Video encoding	T.Aono	Sharp
Video decoding	M.Takahashi	Sharp
Bitrate verification	T.Aono	Sharp
Bitstream verification	M.Takahashi	Sharp
Preparation of test tapes	V. Baroncini	FUB
Conduct of tests	M. Quacchia, V. Baroncini	CSELT, FUB

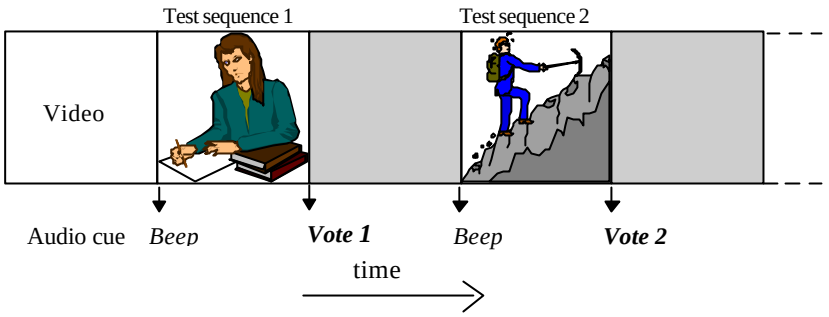
Appendix A: Single Stimulus (SS) Method Instructions

Dear observers,

Thank you for participating in this test.

In the SS tests, a series of test sequences will be displayed on the monitor.

This figure describes what you will see and hear during a SS test session:



Your task is to evaluate the quality of each sequence, by marking one and only one box in the following rating scale:

	☐	10
EXCELLENT	☐	9
	☐	8
GOOD	☐	7
	☐	6
FAIR	☐	5
	☐	4
POOR	☐	3
	☐	2
BAD	☐	1
	☐	0

Your evaluation must reflect your opinion of the global degradation of the whole test sequence. Therefore, vote only after the end of the sequence, and base your evaluation on the entire sequence.

Do not hesitate to rate a sequence either at the top or bottom of the scale, if that is how you believe it should be rated.

A voting form will be distributed before each session. On this form will be a series of rating scales like the one above, one scale for each sequence in the test session. All the scales are numbered. Use scale 1 for the first test sequence, scale 2 for the second one and so on.

During the voting interval for test sequence N, you will hear the audio comment «VOTE N». This will help you know which rating scale to use.

During these tests do not comment on the sequences you have seen or talk with other assessors.

Finally, it is important that you keep your concentration throughout the test session.

Now try this evaluation procedure in a practice session. You will see a series of sequences using the exact same timing as will be used during an actual test session. This will allow you to become familiar with the timing of the test and to practice using the rating scales.

If you have any questions, please ask them now.