Logo

Description automatically generated ISO/IEC JTC 1/SC 29/WG 6 N0084

**ISO/IEC JTC 1/SC 29/WG 6**

**MPEG Audio Coding Convenorship: US**

**Document type:** General

**Title:** MPEG-I Immersive Audio Test and Evaluation Procedures

**Status:** Approved

**Date of document:** 2021-07-16

**Source:** ISO/IEC JTC 1/SC 29/WG 6

**Expected action:** None

**Action due date:** None

**No. of pages:** 27 (including cover page)

**Email of Convenor:** srq @ audioresearchlabs . com

**Committee URL:** https://isotc.iso.org/livelink/livelink/open/jtc1sc29wg6

**INTERNATIONAL ORGANIZATION FOR STANDARDIZATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC1/SC29/WG6, MPEG AUDIO CODING**

**ISO/IEC JTC1/SC29/WG6 N0084**

**July 2021, Virtual**

|  |  |
| --- | --- |
| **Title** | **MPEG-I Immersive Audio Test and Evaluation Procedures** |
| **Source** | **WG 6, MPEG Audio Coding** |
| **Status** | **Approved** |
| **Serial Number** | **20663** |

**Table of Contents**

1 Introduction 3

2 Subjective evaluation 3

2.1 User guidance per Scene-task 3

2.2 Content types 4

2.3 Test design: balanced incomplete block design 4

2.4 Training phase 6

2.5 Testing phase 6

2.6 Test methodology 7

2.7 Post-screening 7

2.8 Test Site preparation for Test 2 8

3 Objective evaluation 8

3.1 Objective metrics 8

3.2 Descriptive input 10

4 Analysis of data and selection of technology 11

4.1 Overview 11

4.2 Figure of merit 11

5 References 13

6 Annex A – Test material 14

7 Annex B – Descriptive input 15

8 Annex C - Statistical analysis of subjective test data 18

8.1 Post-Screening 18

8.2 Pooling of test site scores 18

8.3 Data Analysis 18

9 Annex D – Instructions for test planning 19

9.1 General 19

10 Annex E – Instructions for test subjects 21

10.1 Application background 21

10.2 Training session 21

11 Annex F - AB testing 24

11.1 Testing 24

11.2 How to use the controller in the VR Test + AB testing operations 25

# Introduction

This document provides details on how submissions to the MPEG-I Immersive Audio Call for Proposals (CfP) [1] will be tested and evaluated and how technology from those submissions will be selected as the basis (i.e. Working Draft text and Reference Model 0 source code) for the MPEG-I Immersive Audio specification. The method for evaluation of submissions is based on submitted information and other information, e.g. subjective test scores.

The CfP deliverables shall be separated into two parts, listed below. The timeline for submission of these deliverables is indicated in the CfP.

For subjective evaluation

* Max External audio renderer
* Bit-streams
* Signal latency

Prior to MPEG meeting in which CfP evaluation takes place

* All objective metrics
* Descriptive input

# Subjective evaluation

## User guidance per Scene-task

User guidance is supported in the platform by two means.

* A pre-scene room provides scene-specific information and notes points of attention for the subjects. All scenes shall include such information.
* Within the scene, guiding lines on the floor can be added to guide subjects to points of interest. Interactive scene elements are highlighted.

## Content types

There are three content types in the Scene-tasks that are evaluated in the subjective tests: objects, channels, HOA.

## Test design: balanced incomplete block design

This discussion on statistical test design is based on these references.[[1]](#footnote-2)

A **Latin square** is a *N x N* array filled with *N* different symbols and the property that symbols are exactly once in each row and column.

The typical WG 6 MPEG Audio Coding subjective test is a **complete block** design which has all proposals (proponent CfP submissions, conditions or treatments) in each block. That is each listener evaluates every condition for every test item. A typical example is the Latin square design, which is a *N x N* array filled with *N* different symbols and the property that symbols are exactly once in each row and column.

However, if the number of proposals to be evaluated by one lister would exceed the available time allocated for a test, then only a subset can be evaluated. This is called **incomplete block design**.

A design is called disconnected if there are groups of proposals that are never compared. This should be avoided, because it makes comparison between proposals in different groups impossible.

In a **balanced incomplete block design** (BIBD) all pairs of proposals (treatments) occur equally often in each block. The number of occurrences is denoted *lambda*. The precision (variance) of differences between blocks is the same for all pairs. The following table relates the usual wording from statistics textbooks to the parameters (and values) as used in WG 6 (MPEG):

|  |  |  |  |
| --- | --- | --- | --- |
| **Block Design (textbook)** | | **MPEG** | |
|  | |  | |
| Number of treatments | g | Number of Proposals | P |
| Number of Blocks | b | Number\_of\_listener \* number\_of\_sessions per\_listener |  |
| Number of units per block (k<g) | k | Number of scorings per listener and trial | C |
| Number of replicates per treatment | r | Number of scorings by one listener for each proposal |  |
| Total number of units | N |  |  |

The following must be fulfilled: N = b\*k = g\* r

All values in this equation must be integers. Therefore, the selection of r is not free.

As long as k<g there is always at least the BIBD where all combination g\_over\_k are used. This is called an **unreduced balanced incomplete block**. Typically in this case the number of blocks is too large. In context of this subjective test, we would need either too many listeners or too many sessions per listener.

To **reduce** the amount of testing the following way has to be used:

Each proposal occurs in exactly r blocks. If we have selected one proposal there are k-1 other proposals in each block making a total of r\*(k-1) available units. The remaining g-1 proposals must be divided evenly among them to make the design balanced.

Lambda= r\*(k-1)/(g-1) must be a whole number.

This condition is necessary but not sufficient. There is an R function in the library crossdes which can be used to find a proper design.

The following table shows some results for g=14 (14 proposals) and different number of scores per block:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **k** | **Lambda**  **r \* (k-1)/g-1)** | **Min. Lambda** | **r for min. lambda** | **N r\*g** | **b  N/k** | **# of listener per site (min.)** | **Remark** |
| 2 | r \* 1/13 | 1 | 13 | 182 | 91 | 7 | Pair test |

Based on these numbers the estimated duration of the test is (two listeners interleaved):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test Methodology** | **Blocks per listener (trials)** | **Time per listener [h]** | **Worksdays per listener** | **Workdays per site (with min. # of listeners)** |
| AB Comparison | 13 | 18.2 | 4.6 | 15.9 |

A strict BIBD design requires a large number of listeners and furthermore, an exact number of listeners. In the MPEG test, listeners might decide to not participate, or might be removed by post-screening. Hence, in this test the exact requirements for BIBD must be relaxed.

The test must be sufficiently balanced across all test sites. Each listener will evaluate a subset of scene-task pairs, with a unique randomization. The test files for every listener will be created by the test coordinator based on the script provided by Ericsson/Philips/FhG-IDMT.

## Training phase

Training is mandatory for all test subjects, and consists of the following two components.

**Familiarization with acoustic effects**

Subjects shall watch the following YouTube videos that are illustrative of certain acoustic effects and how they manifest themselves aurally.

Steam Audio: Partial Occlusion  
<https://youtu.be/9lywYAGSgdc>  
  
Steam Audio: Propagation  
<https://www.youtube.com/watch?v=y2vqK6pDYa0>  
  
Phonon Soundflow Beat Demo:  
<https://www.youtube.com/watch?v=q5MC1FJ1pec>  
  
Phonon Loft Architecture Demo  
<https://www.youtube.com/watch?v=p6MlMn5g0FM>

**Familiarization with test environment and test content**

Subjects shall use a modified version of the CfP subjective test for training purposes. The training task shall include:

* All 14 Test 1 Scene-tasks for Test 1 training
* All 7 Test 2 Scene-tasks for Test 2 training
* All 4 Test 3 Scene-tasks for Test 3 training
* Each Scene-task shall present all submissions in at least one AB assessment panels (e.g. 14 submissions is 7 panels). The ordering of submissions shall be randomized over the set of Scene-tasks. There is only one training listener file. However, the subject can “skip ahead” through remaining panels if the task in the scene is well understood.

Subjects training shall use training time in VR/AR scenes to get used to the HMD, controllers, moving around in VR/AR, virtual interfaces and the Scene-tasks and proponent renderers. This shall be done with a special training listener file that includes all proponent renderers for all Scene-tasks. The renderers order shall be in a randomized order for each Scene-task.

The training session should *not* take place immediately prior to testing.

## Testing phase

Tests are defined in [1], and summarized here:

Test 1: VR with Objects, Channels, 3DoF HOA

Test 2: AR with Objects, Channels

Test 3: VR with 6DoF HOA (interior/exterior, multi-point), Objects, Channels

Test material for each test is given in Annex A.

Proponents are encouraged to submit technology to be evaluated in all three tests in the CfP (Test 1, Test 2, Test 3). Proponents are not required to submit technology for all tests and could submit for only one or two tests. However, submissions to a given test shall be evaluated for all Scene-tasks and associated audio signal content types in that test.

Test Sites can choose whether they perform one test or more. It is expected that there is sufficient interest for all tests to get results from multiple Test Sites per test.

Noise level in the Test Site subjective test room should be low since open-back headphones are used in the test. If possible, the PC running the AEP should be moved to outside the listening room to minimize PC noise. The recommended minimum play area size is 3m by 3m and the maximum size is 4.5m by 4.5m.

## Test methodology

The MPEG-I Immersive Audio CfP listening tests are performed using the A-B methodology [2]. See additional details in Section 4.

## Post-screening

### Introduction

Post-screening is applied to subject scores. There are two types of post-screening. The first is based on subject scoring, the second is based on the subject’s task conformance.

### Subject scoring

Post-screening based on subject scoring is achieved by including 4 self-comparison trials in each listener file, where A and B are the same renderer. Subject’s results are retained if at least 3 self-comparison scores are within the score range -0.5 to 0.5 inclusive *and* all 4 self-comparison scores are within the score range -1.5 to 1.5, inclusive.

### Subject task conformance

* Scene-task creators shall define a “check list” of conformance criterion for their Scene-task. This can be as simple as: did subject go behind an occluding object?
* All log files from all test sites will be retained and made available for review after the subjective tests are concluded. These could be used for subsequent analysis of subject task conformance. It is anticipated that subjects would be excluded only for extreme conformance negligence.

## Test Site preparation for Test 2

### AR listening Test Site requirements

The size of the AR lab needs to be big enough to accommodate the AR scenes. The largest scene selected in the CfP Test 2 is the Singer in Your Lab (small), which requires a room that at least has a height of 2.5 meters and a length and a width of at least 3 meters and 3 meters, respectively. The AR Test Sites are encouraged to use simple rooms that have not been acoustically optimized. For example, remove unnecessary furniture.

### Measurement of LSDF elements

An Listener Space Description Format (LSDF) file [3] is used to describe the listening space for Test 2. RT60 and DDR values for the <AcousticEnvironment> defined in the Test Site LSDF file may be obtained from measurements or room simulations. Using measurements is encouraged over room simulations. The RT60 and DDR values should be determined from the same measurement process (see [4] for details). Acoustic material properties for the room mesh faces may be obtained from a material database or by making an informed estimate. The <Mesh> element properties in the LSDF file is obtained by manually measuring the listening space.

### LSDF submission timeline

LSDF files are shared by Test Sites for cross-checking purposes. The LSDF files should be verified and ready for all the Test Sites before the start of the listening tests. The following timeline shall be followed.

|  |  |  |  |
| --- | --- | --- | --- |
|  | What | When | Details |
| 1. | LSDF submission by Test Sites | 8 weeks before the start of subjective testing | Test Sites share LSDF file, picture of test room, description of measurements for RT60 etc. and a description of placement of anchors. Sharing done by CFP Test Administrator via MPEG GitLab. |
| 2. | Cross-check of LSDFs | 4 weeks before the test date | At least one company needs to cross-check each of the LSDF files. Sufficient time should be allocated to allow for a few cross-check and fixing iterations. |
| 3. | Distribution | 2 weeks before the test date | Done by the CfP Test Administrator. |

# Objective evaluation

## Objective metrics

The CfP stipulates that the following objective metrics shall be reported by proponents:

* computational complexity,
* bitrate,
* latency (motion to sound latency as well as signal throughput latency),
* memory usage.

***Computational complexity***

A computational complexity estimate of the CfP renderer implementations shall be reported by the CfP proponents using the “*plugin complexity self-comparison*” method. This comparative complexity estimation method is designed to identify a significant difference in CfP solution computational complexity. The corresponding profiler instrument for this method is available on [GitLab](http://mpegx.int-evry.fr/software/MPEG/ImmersiveAudio/audio_evaluation_platform/-/commits/feature/complexity-measurements)[[2]](#footnote-3).

The “*plugin complexity self-comparison*” method obtains an estimate of computational complexity of CfP renderer implementation complexity (i.e., plugin ) as follows:

where P is a workload measurement of the AEP system hosting one instance of a proponent plugin and PREF is a workload measurement of the AEP system hosting two instances of the same plugin. Complexity shall be expressed as an accumulated processor time of the Max process. Complexity measurements shall be reported for the CfP Test 1. The detailed instructions on how to perform the measurements using the corresponding tools are available on GitLab.

Complexity measurements shall be performed for predetermined user pose-traces (potentially including simulated user-interaction) within the corresponding user play-area. Time resolution of the user pose data shall be 10 Hz. The measurement duration shall be 10 minutes per each test scene. The predetermined user pose-traces shall be recorded and made available by each Test Site.

All Test Sites shall conduct the complexity tests for all proponent renderers on one machine and report this in a contribution to the WG 6 meeting at which the CfP submissions evaluation is done. The final complexity used to select technology will be the median of the distribution of results for each proponent submission.

***Latencies***

Motion to sound latency and signal throughput latency shall be expressed in PCM samples at a sample-rate of 48000 Hz. Motion to sound latency shall be provided for all subjective listening tests in which the proposal participates. Latency shall be measured without any modeling of sound propagation time (i.e. aparams = “nodistance”). This shall be computed using “pencil and paper” methods based on the algorithms.

***Memory usage***

Proponents shall estimate memory usage and describe the method used.

The memory usage shall be estimated and reported by the proponent for a pre-determined scene. They shall express the highest memory load during each such scene and pose-trace combination.

***Bitrate***

Bit-streams provided for the CfP shall be defined such that the random access delay is no more than 1 second. This means that in a streaming case an MPEG-I Audio decoder/renderer tunes in to an ongoing stream, it should take no longer than 1 second of scene progression time for the decoder to have received all data needed for a full quality rendering.

For the CfP each proposal shall use one of two concepts to achieve this:

1. Traditional independent bit-stream solution that repeats all data needed for a full quality rendering at least every second.
2. Interactive server-client solution where a client requests and receives an init package extending the regular bit-stream with all remaining data for a full quality rendering.

Bitrates for all CfP proposals shall be determined and reported for all provided 6DoF metadata bitstreams. Bitrates shall be calculated according to the following equations, depending on which concept is used by the proposal:

A:

B:

where duration is measured in seconds and that Kb means kilobits.

All scenes have a duration, reflected in the duration of their continuous signal WAV files, or shall be assigned a default duration of 3 minutes in case there are no continuous signals.

## Descriptive input

Not all MPEG-I Audio CfP requirements are tested in a subjective test (sub-scening, social VR interface, extensibility, etc). Therefore, solutions to these and all other requirements shall be described in sufficient detail so that audio experts understand how the requirement is met and can make a technically qualitative judgement about it.

The reporting points listed in the table in Annex B shall be used by proponents to prepare descriptive input.

# Analysis of data and selection of technology

## Overview

The selection of a proponent technology as RM0 shall be made by considering subjective performance, objective performance and descriptive information. The subjective performance is of greatest significance.

## Figure of merit

For the Immersive Audio CfP evaluation, the subjective quality is taken as the Figure of Merit (FoM). Specifically, the FoM for a condition for a given test (Test 1, Test 2, Test 3) shall be the grand mean for that condition, as averaged over all listeners from all Test Sites (as appropriate for pooling) and all Scene-Tasks. This mean is computed using the Thurstone Case V analysis (see Annex C). Additionally, the grand mean is associated with a 95% confidence interval on the mean. The proposal with the highest FoM will be designated as the “*winner*”, but see details in the remainder of this section.

Proposals may target different “*operating points*” based on rendering complexity and bitrate.

If there is another proposal that has a “*significantly different operating point”* than the FoM winner, then one additional “*category winner*” will be designated and its technology will be merged/combined into the “*winner*” solution.

The “*significantly different operating point”* (referred to as the category) is defined based on the complexity and bitrate of the “*winner*” (highest FoM) proposal as follows:

* low complexity category
* low bitrate category

The values for and are obtained as described in Section 3.

Only one “*significantly different operating point”* is considered. The complexity category is preferred and considered first. If there is a “*category winner*” in the complexity category, then the bitrate category is disregarded. If there is no “*category winner*” in the complexity category, then the bitrate category is taken into consideration.

More specifically, for Test 1:

* Proposals will have a range of reported complexities. The threshold defined above of these complexity numbers splits the proposals into “low” and “high” complexity categories. One best proposal per each “low” and “high” complexity category is chosen based on FoM. Let us assume that the “winner” proposal is in the “high” complexity category. If the complexity of the winner technology in the “low” category has complexity less or equal to 20% of the “winner” technology in the “high” complexity category, then the unique modules of this “category winner” system that are responsible for the complexity reduction will be merged/combined with the first (highest FoM) “winner” system.

If this procedure results in no “category winner” for the complexity category, then the same procedure is applied to the bitrate category with the corresponding bitrate ratio threshold of 5/12, i.e., the unique modules of this “category winner” system that are responsible for the bitrate reduction will be merged/combined with the first (highest FoM) “winner” system. For bitrate category the same criterion and procedure are applied as for complexity category.

If no proposal satisfies this condition, then the “category winner” will not be designated.

If the proponent condition with the highest FoM in a category has a confidence interval that overlaps with the confidence interval of other proponent conditions in that test or category, the Audio working group shall review the objective metrics, descriptive input and Scene-task dependent performance of these proponent conditions. The winner shall be chosen as the condition that has best overall performance. In case proposals with overlapping confidence intervals are complementary in any way it shall be explored whether a merge provides a better and more flexible standard than each proposal by itself.

**Merging of winning technologies**

The Test 1 “winner” forms RM0 base.

The proposal with the highest FoM among the corresponding category (if any) will be designated as the “*category winner*”. The “*category winner*” technology merges into Test 1 “*winner*” solution.

* Test 1 "category winner" merges "unique" modules into Test 1 RM0 base

For Tests 2 and Test 3: No “*category winner*” is designated.

* Test 2 "winner" merges "unique" modules into Test 1 RM0 base
* Test 3 "winner" merges "unique" modules into Test 1 RM0 base

All Test 1, 2 and 3 "winners" (and Test 1 "category winner") proposals shall demonstrate “*good performance*”. At the option of WG 6 experts, a subjective test using one of the BS.2132 [5] methodologies may be conducted on the WD technology. If the outcome does not demonstrate a subjective quality of “good” or better, then if may be decided to not proceed with further development and standardization of the WD technology.

# References

1. N0056, MPEG-I Immersive Audio Call for Proposals
2. ITU-R Recommendation BS.1284-2, General methods for the subjective assessment of sound quality, 2019, Intern. Telecom Union, Geneva, Switzerland (specifically, A-B Comparison)
3. N0055, MPEG-I Immersive Audio Augmented Reality Listener Space Description Format
4. N0083, Clarification Document?
5. ITU-R Recommendation BS.2132, Method for the Subjective Quality Assessment of Audible Differences of Sound Systems Using Multiple Stimuli Without a Given Reference, 2019, Intern. Telecom Union, Geneva, Switzerland.
6. N0028 MPEG-I Immersive Audio Architecture and Requirements
7. Perez-Ortiz and Mantiuk “A practical guide and software for analysing pairwise comparison experiments” (https://arxiv.org/abs/1712.03686

# Annex A – Test material

Submissions to the CfP shall use

* Diffuse Field Equalized Fabian HRTF (supplied with AEP)
* Apply no Headphone Equalization

Continuous (i.e. streaming) audio signals associated with all test material shall be encoded and decoded using the MPEG-H 3D Audio codec operating at the following bitrates:

|  |  |
| --- | --- |
| **Signal** | **Coding Bitrate** |
| Monophonic object: | 64 kb/s |
| Stereo object | 128 kb/s |
| 22.2 “channel source” or set of objects | 768 kb/s |
| Test 1, Order 1 HOA signal | 256 kb/s |
| Test 1, Order 4 HOA signal | 512 kb/s |
| Test 3, Order 1 HOA signal | 256 kb/s |
| Test 3, Order 3 HOA signal | 768 kb/s |
| Test 3, Order 4 HOA signal | 1024 kb/s |

Proponents may obtain test material via FTP. FTP site (URL), username and password will be sent to proponent after proponent registration (See CfP [1]).

**Test 1**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Attribute** | **Item Name** | **Provider** |
| 1 | Size/shape of a sound source | Park | FhG-IIS |
| 2 | Directivity | SingerInTheLab | FhG-IIS |
| 3 | Occlusion (by obstacles; partially translucent & hard) | Park | FhG-IIS |
| 4 | Diffraction (around corners …) | SimpleMaze | Dolby |
| 5 | Coupling/sound propagation between adjacent rooms | Hospital | Philips |
| 6 | Early reflections (from walls, rocks, …) | DowntownBus | FhG-IIS |
| 7 | Late reverberation | VirtualBasketball | FhG-IIS |
| 8 | Transition between different environments | Cathedral | FhG-IIS |
| 9 | Distance attenuation and far field effects (behavior in a distance) | Beach | Ericsson |
| 10 | Doppler | ParkingLot | FhG-IIS |
| 11 | Interaction (M) + L1 + 6DoF localization (N) | Battle | Philips |
| 12 | Interaction (M) + L3 + 6DoF localization (N)  (grab radio and go to “tent” area) | Canyon | FhG-IIS |
| 13 | Teleport transitions between very different acoustic environments | Recreation | Philips |
| 14 | 3DoF HOA | OutsideHOA | Qualcomm |

Table A-1 -- Material for Test 1

**Test 2**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Attribute** | **Item Name** | **Provider** |
| 1 | Directivity | AR\_BMW | FhG-IIS |
| 2 | Occlusion (by obstacles; soft/partially translucent & hard) | AR\_VirtualPartition | Nokia |
| 3 | Diffraction (around corners) and Coupling/sound propagation between adjacent rooms | AR\_Portal | FhG-IIS |
| 4 | Plausibility of Early and Late reflections | AR\_Portal | FhG-IIS |
| 5 | Distance attenuation and far field effects (behavior in a distance) | AR\_SingerInYourLab (small) | Nokia |
| 6 | Interaction (M) + L1 + 6DoF localization (N) | AR\_GigAdvertisement | Nokia |
| 7 | Interaction (M) + L3 + 6DoF localization (N) | AR\_HomeConcert | FhG-IIS |

Table A-2 -- Material for Test 2

**Test 3**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Attribute** | **Item** | **Provider** |
| 1 | Overall User Experience | MultizoneMusicHOA | Nokia |
| 2 | Overall User Experience | WindyForest | Ericsson |
| 3 | Overall User Experience | Borchestra | Qualcomm |
| 4 | Overall User Experience | LockedRoom | Zylia |

Table A-3 -- Material for Test 3

# Annex B – Descriptive input

Proponent must submit a description of how their technology fulfills certain MPEG-I Immersive Audio requirements, indicated in Table B-1, below. Those requirements have the entry “Description” under the heading “Methodology.” The table lists all requirements found in [6] by their number and gives a short description of the requirement. The additional columns indicate an assessment methodology for the requirement and any relevant comments concerning the methodology.

The Methodologies are:

|  |  |
| --- | --- |
| Subjective | Assess via subjective test, described in Section 2, above. |
| Objective | Proponent-conducted objective test, described in Section 3, above. |
| Description | Proponent description of how requirement is satisfied, as described below. |

For requirements whose assessment methodology is *Description*, the proponent submitted documentation must contain a description of how the proposed technology satisfies the requirement. This description must be sufficient for Audio subgroup experts to check on their own that the requirement is met.

|  |  |  |  |
| --- | --- | --- | --- |
| **Requirement Number**  (from [6]) | **Requirement** | **Methodology** | **Description** |
| 1 | Perceived experience | Subjective | Performance is assessed via listening tests. |
| 2 | Efficient representation | Objective | State complexity and metadata bit-rate. |
| 2.1 | MPEG-H 3D Audio coding | Description | State how MPEG-H 3D Audio LC profile bitstream would be represented in a unified MPEG-I bitstream. |
| 2.2 | DRC | Description | Proponent describes how this is done. |
| 2.3 | MPEG-I metadata | Description | Proponent describes metadata. |
| 2.4 | Multiple streams | Description | Proponent describes how this is done. |
| 3 | Consistent experience | Subjective | Performance is assessed via subjective tests with HMD with headphone listening. |
| 3.1 | Consistent visual and audio | Subjective and Description. | Performance is assessed via subjective tests with HMD with headphone listening.  Describe how match with visual elements is achieved in time and space. |
| 3.2 | Non-diegetic audio elements | Description | Proponent describes how this is done. |
| 3.3 | Earcons | Description | Proponent describes how this is done. |
| 4, 4.1 | Division into sub-scenes | Description | Proponent describes how this is done.  Describe sub-setting process and algorithm (if any).  Describe any associated interfaces and metadata. |
| 5 | Metadata describing scene | Description | Describe MPEG-I metadata for scene. |
| 6, 7 | Controlling or restricting audio scene or parameters | Description | Describe MPEG-I metadata or interfaces to support restrictions and how the restrictions are performed. |
| 8 | Random access | Objective and Description | Describe the bitstream structure needed to support random access in time and space, and report bitrate for the test scenes coded with 1s random access intervals.  Describe how both fully independent bitstream mode as well as interactive bitstream exchange mode with server is supported. |
| 9, 10 | Rendering for jumps, accelerated movement and zoom | Description | Describe the metadata that supports this functionality.  Describe how this is done and what the user would perceive. |
| 11-13, 15 | Rendering requirements | Subjective and Description | Describe the metadata that supports this functionality.  Performance is assessed via HMD and headphone listening tests that stress the requirements of the renderer. |
| 14 | Locally captured audio | Description | Describe how local voice can be provided to the renderer and how it is rendered.  Describe how other sources of local audio are supported (e.g. ambient sound). |
| 16 - 18 | Interfaces | Description | Describe specified interfaces and how they support this functionality. |
| 18, 19 | Personalization of HRTF, HP EQ | Description | Describe the specified HRTF and HP EQ interfaces and how they would support the desired personalization. |
| 20 | 6DoF with headphones | Subjective | Performance is assessed via subjective tests with HMD and headphone listening. |
| 21 | 6DoF with loudspeakers | Description | Describe algorithms used with head tracker and loudspeaker listening. |
| 22 | Static user with joystick | Description | Describe specified interfaces and how they support this functionality. |
| 23 | 6DoF with both headphones and loudspeakers | Description | Describe how technology fulfilling requirements 20 and 21 can be combined into a single presentation |
| 24, 24.4 | Social VR | Description | Describe specified interfaces for remote user media (speech/audio) and metadata (6DoF position and orientation) and how this is supported in proponent architecture and technology. |
| 24.1, 24.2 | Social VR | Objective and Description | Measure algorithmic latency subject to stated assumptions |
| 24.3 | Synchronization of A and V | Description | Describe how proponent metadata and technology would assure synchronization. |
| 25-27 | Interoperability | Description | Describe how this would occur, with reference to envisioned bitstream syntax or decoder functionality.  Describe the envisioned user experience. |

# Annex C - Statistical analysis of subjective test data

## Post-Screening

Post-screening of subjects participating in the CfP listening tests shall be done as described in Section 2.6.

## Pooling of test site scores

Since factors are balanced only across the entire set of test sites and listeners, all listener scores remaining after post-screening are pooled together.

## Data Analysis

The AB score data analysis is done using the Thurstone V analysis to convert AB preferences to an ordering of preference with 95% Confidence Intervals. The analysis is explained in greater detail in the paper by Maria Perez-Ortiz and Rafal K. Mantiuk, “A practical guide and software for analysing pairwise comparison experiments”, available at <https://arxiv.org/abs/1712.03686> [7]. Perez-Ortiz and Mantiuk have created Matlab code to implement the Thurstone V analysis, and this Matlab code is used for the analysis of the CfP AB scores.

The AB preference data is reduced to “win-loss” data, where if A is Better, Slightly Better or Much Better, then A is assigned a “win” (score of 1) versus B. Conversely, A can be assigned a “loss” (score of 0) if B scores Better, Slightly Better or Much Better. If the score is The Same (in the interval -0.5 to +0.5, inclusive), then both A and B are awarded a “win” score of 1/2.

To get the data analysis Matlab code and understand how to use it, refer to the ReadMe.md file in the AEP DataAnalysis folder on the MPEG GitLab.

# Annex D – Instructions for test planning

## General

In comparison to previous MPEG Audio test efforts, MPEG-I 6DoF Audio subjective quality evaluations are more demanding since they not only involve the subject’s auditory perception as a primary sense used for evaluation, but also the subject’s sense for visual perception, body motion and the multi-modal integration of those. Generally, it is the intention to

* Limit testing time and thus achieve a high test efficiency
* Limit stress on test subjects (given the demanding multi-modal test environment) and thus improve general grading quality
* Reduce cognitive complexity of test task and in this way also improve grading consistency

If test site Local Test Adminstrators wish to communicate with test subjects, e.g. to offer guidance during a test, they can speak via a microphone connected to the Scarlett audio interface microphone input, and put that interface in “direct monitor” mode so that the microphone signal is directly mixed into the headphone signal.

**Training**

The subject training session is essential to reduce cognitive overload, to have stable categories and scores and to get familiar with all scenes and quality levels. The duration of the training session depends on the number of different scenes and systems under test. Sufficient time has to be planned for each subject. The training should be done one or several days before the actual test. All virtual scenes are presented for training purposes.

**Testing**

The test instructions for the test subjects should be read by the subjects and also be read to the subject by the Test Supervisor.

**Test Subjects**

A naïve subject is a listener who did get training on the test methodology but does not have a long experience with this kind of test, whereas an expert subject is a listener who has done many audio attribute testing over months if not years. The test panel should consist of sufficiently many naïve and expert subjects such that it is possible to do separate statistical analysis for both groups. As the level of experience with VR/AR may influence the test load on the subject, it should also be noted whether the subject is a naïve or experienced VR/AR user.

**Test Environment**

The test environment needs to be reported regarding equipment use, size of the room and acoustic properties. It is recommended, that the test subjects have a play area at least an area of 3m by 3m and not larger than 4.5m by 4.5m. Regarding acoustics of the listening room, it is not required to fulfill the requirements of standards like ITU-R BS.1116. The acoustic properties (dry or more reverberant room) probably have some influence on the immersiveness of the reproduction. MPEG Audio would like to see test results from both kinds of rooms.

**Testing Recommendations**

* Limit test session time to 30 minutes, with breaks between test sessions.
* Training must must permit subject to become acquainted with all scene-tasks in the test. Also, to accommodate their natural curiosity, subjects should be permitted to fully explore one or more scene-tasks in a leisurely way, without regard to designated scene-task.

# Annex E – Instructions for test subjects

## Application background

The purpose of the provided Virtual Reality (VR) environment is for personal entertainment, information, media consumption, gaming or other purposes in your work or home environment. The environment provides both visuals and acoustics and you should feel immersed as much as possible. Please consider how comfortable you would be with using such an environment far longer than in this test.

## Training session

You must have training prior to the actual testing. The purpose of training is for you, the subject, to become familiar with the assessment process, the virtual scenes and the different aspects of audio quality. This training does not take place immediately prior to testing. All virtual scenes are presented for training purposes.

Firstly, you should learn to use the physical test equipment and practice your ability to move and interact with the virtual scenes, including:

* Getting the correct fit for the HMD and headphones
* Understanding how to use the controller(s)
* Understanding how to start the test and to proceed through the different test scenes
* Understanding how to navigate within a specific test scene:
* User movement (yaw, pitch, roll)
* User movement (x, y, z), including teleportation, and touchpad walking to extend (x,y,z) navigation range
* Understanding how to bring up the scoring response panel and how to operate the various controls on the panel to switch between the different test conditions and rate those conditions.

Lastly, it is important to realize that specific navigation may be necessary to assess certain scene audio characteristics, including:

Table E-1 – Perceived Quality Features (QF) for real-time auditory virtual environments[[3]](#footnote-4)

|  |  |  |
| --- | --- | --- |
| **Quality Feature/Attribute** | **Description** | **User Movement or Navigation** |
| Localization accuracy | Discrepancy between the desired and the perceived location of the auditory event in the 3-D space | head & body movement |
| Timbre | Plausible timbre, no unwanted sound colouration | listen to familiar sound sources like human voice |
| Loudness balance | Plausible balance of loudness between different sound sources | listen to the balance in different distances from a sound source |
| Directivity | The way a sound source emits sound to different directions influencing the timbre | navigate around a sound object |
| Reverberance | Plausible reverberance (room perception), caused by echoes/reflections produced by a room when excited by sound sources | navigate to various locations in scene |
| Occlusion | Change in timbre due to obstacles between the sound source and the listener | navigate around and behind sound occluding element |
| Source width | Perceived spatial extent of a sound source | navigate along element with head movement (when close) or by physical walking (when distant) |
| Dynamic accuracy | A sound source is moving smoothly and without perceived time delay regarding the interactive input movement. | listen to moving sound sources or move as a listener |
| Audibility of Artifacts | Prevention of audibility of artifacts (errors) which are unwanted auditory events like clicks, … |  |
| Scene transition | Change in sound characteristics due to transition between different scene parts | navigate through doorways to reach different areas |

Additional “what to listen for” instructions are provided to you for each test scene, and must be read prior to exploring that scene. This may include scene-specific interaction or navigation instructions on “how to move & interact” and also may include additional cues inside the scene e.g. red floor markings indicating a recommended navigation path.

If you are unsure about the proper testing procedure, the Test Supervisor can discuss with you how certain locations and elements in the training scene to exemplify significant aspects to be assessed.

# Annex F - AB testing

## Testing

The duration of testing may vary. The Test Supervisor will manage testing so that you do not go for too long (e.g. 30 minutes) without a break.

From the training phase, you should already be familiar with how to use the controller to operate the scoring panel and navigate via teleportation or touchpad walking. However, as a reminder, the test gear is operated as summarized at the end of these instructions.

The Test Supervisor will verbally explain to you how to use the controller to navigate via virtual teleportation and touchpad walking and to bring up the scoring panel for switching between the different test conditions and recording your rating. The test condition that is currently active and that is to be rated is indicated by a high-lighted slider.

You should actively navigate and interact with the scene by user movement in terms of orientation (yaw, pitch and roll) or translational movement (x, y, z) - either performing near-by movement by physical walking or far-off translation using virtual teleportation or touchpad walking (if enabled in the scene). You should try to navigate within the “play area” by walking whenever possible, and use teleportation to move to a new “play area” location which is then further explored via walking. For each new scene, try to start in the middle of the available “play area”.

You are free to explore the scene as you wish. However, scene specific guidance information on your required movement and interaction types will be displayed upfront prior to exploring each of the scenes and additional guidance information within each scene will be provided. It is mandatory in any case to reenact what is described in the guidance information.

To issue your rating of each scene, please bring up the scoring panel and evaluate the **overall Quality of Experience** of the audio reproduction within that scene in the current condition. By adjusting the slider’s value on the scoring panel, you indicate your personal assessment of the comparative quality of A versus B on a scale ranging from “Much Worse” than to “Much Better” than.

Scores for a system under test must encompass all aspects of the perceived audio experience in the scene, including all Quality Features described in Table E-1.

At the end of the test please report to Test Supervisor any observations concerning test procedure, scenes, or audio quality issues in free form text.

## How to use the controller in the VR Test + AB testing operations

1. Read through the functionality for the controls, then put the HMD on and familiarize yourself with the VR environment.
   1. The instructions shown in Figure 1 for controller usage are also provided inside the ‘\_Main’ scene for you to refer too. These will *not* be shown again in subsequent scenes throughout the test, so make sure you can easily control what you would like to do so as not cause a distraction during evaluation.

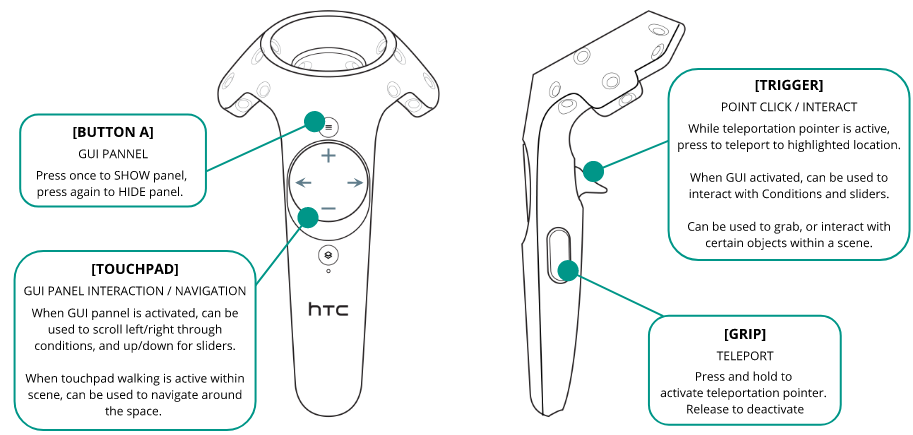


Figure 1 - HTV Vive Control Usage

* 1. Teleportation within the AEP uses a point and click, blink system (Figure 2). Holding down the [GRIP] button will activate a pointer, displaying a Bézier curve and target location. The Bézier curve is used to restrict the teleportation to a maximum distance. Holding down the [GRIP] button and pulling the [TRIGGER] will teleport the user to the target location. A small fade-out/in (blink) of visuals is present, and no matter how far the distance, teleportation time is always 5 ms. Please try to avoid teleportation, whenever it is possible.

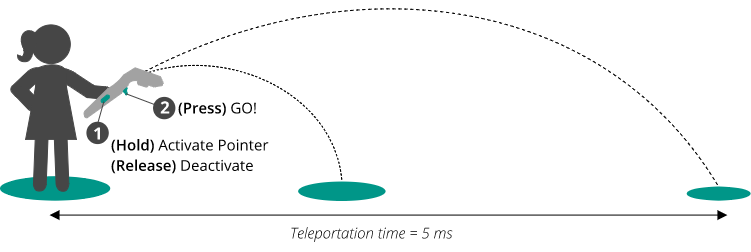
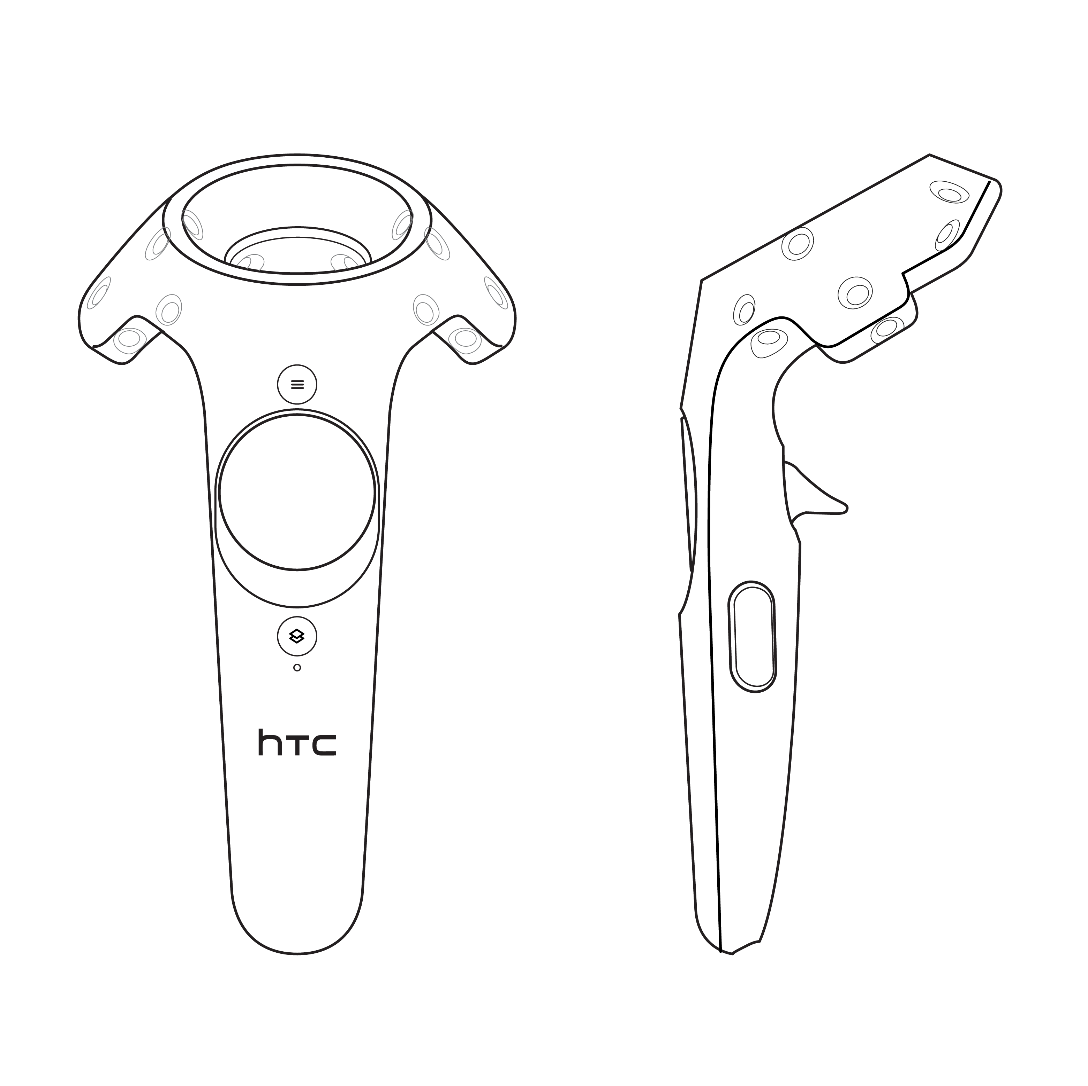


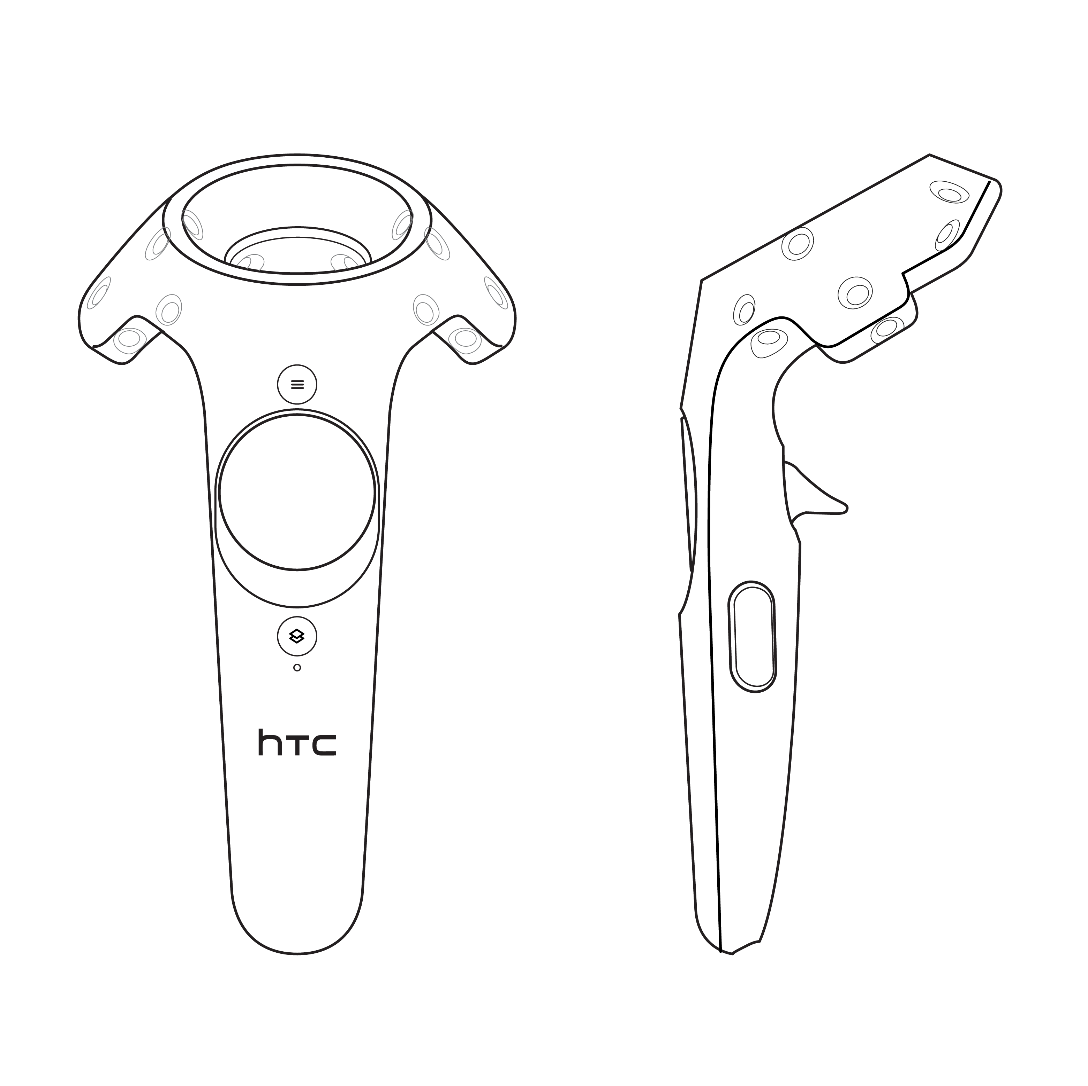
Figure 2 - Teleportation Usage

* 1. Some scenes make use of an alternative way of moving within a scene where the user can glide smoothly in any direction by pointing with a hand controller and using the touchpad to glide forward, backwards, left or right relative to the pointing direction (see Figure 3). Touching the touchpad closer to the center will make the movement slow whereas a touch closer to the edge will result in a faster movement. This can be useful e.g. when the user wants to move over larger distances while evaluating rendering aspects that vary with the position. This functionality is disabled whenever the UI panel, as described below, is brought up.



Touching the touchpad at the top middle moves the user in the direction that the controller is pointed.

Touching any other part of the touchpad will move the user to the right, left or backwards relative to the direction of the controller.



Touching the touchpad at the top middle moves the user in the direction that the controller is pointed.

Touching any other part of the touchpad will move the user to the right, left or backwards relative to the direction of the controller.

**Figure 3 –Touchpad Walking Usage**

* 1. Pressing [BUTTON A] on the controller will bring up the scoring Panel, this is the main interface to use throughout the evaluation. The Panel will always instantiate in the user’s field of view, after which, it will remain static at this position. If you wish to reposition the panel, press [BUTTON A] to hide the panel, then again to re-instantiate it.

At the same time the panel appears, a laser pointer is also activated. The laser point can be used to interact with the buttons on the panel (see Figure 4). Mouse over a button using the laser pointer will highlight the item, pressing the [TRIGGER] will ‘push’ the button. As shown above in Figure 1, the touchpad can be used to control conditions and sliders, too.

A-B Operations:

|  |  |
| --- | --- |
| * ‘**A**’, ‘**B**’ | Select conditions A or B. |
| * *Slider* | Indicate a relative rating for the two conditions. |
| * ‘**Next**’ | Continue to the next evaluation scene. |

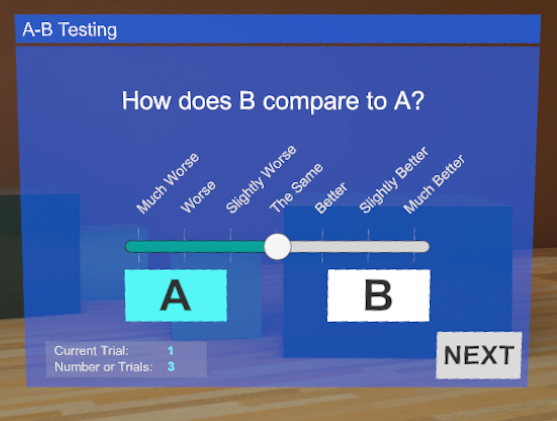
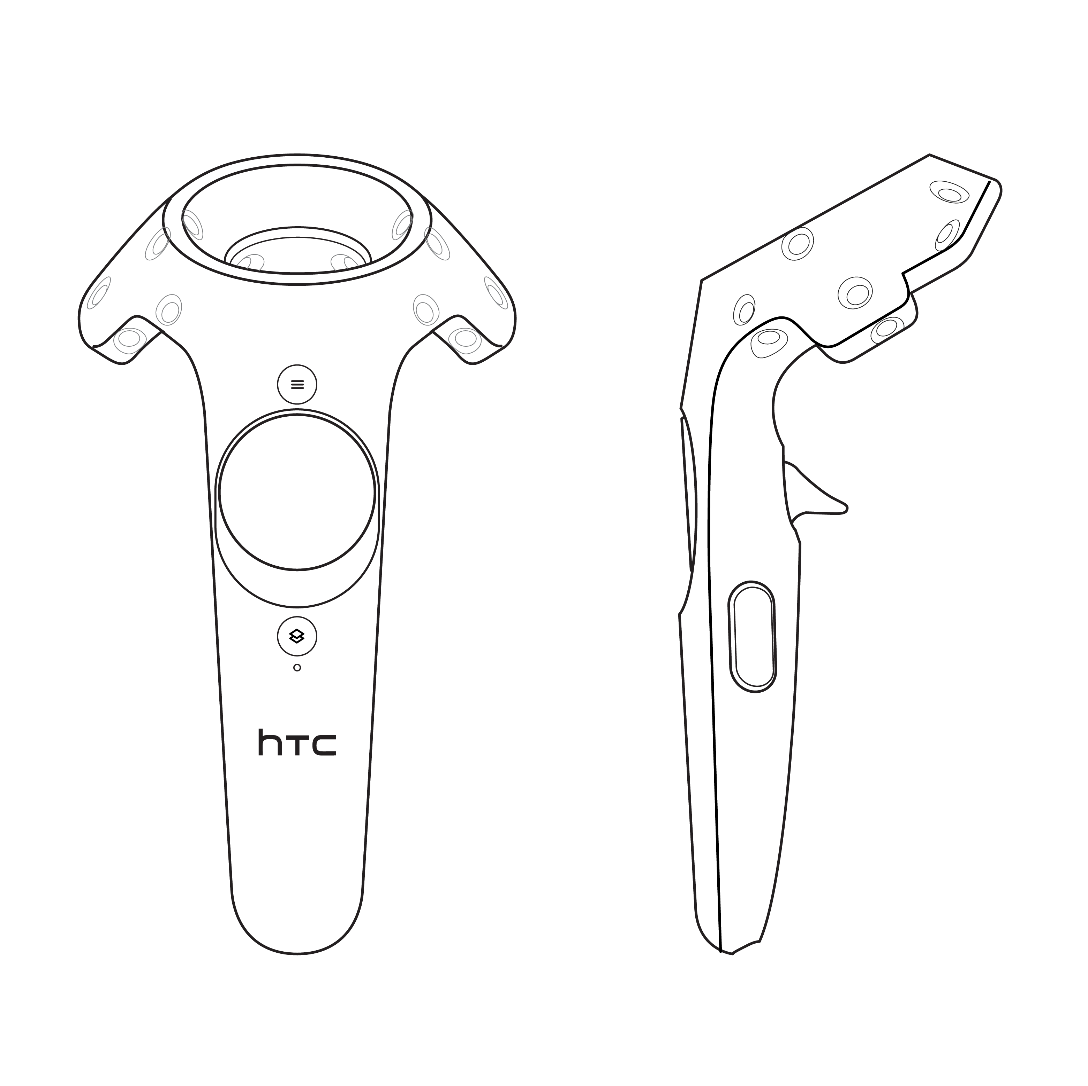


Figure 4 - Test Panel Interactions

1. Once you are accustomed to the controls, teleporting, touchpad walking and operating the scoring panel, navigate to the main menu and press `Begin Test'. This will then load the first evaluation scene.
2. Upon entering a new scene, you will be presented with an empty room that has two images. These images are the pre-scene instructions and a topology illustration of points of interests within the scene. Please take the time to read the instructions carefully and memorize the points of interest. Once you are ready to begin evaluating the content, a red capsule is located in the room next to the right of your position. Take the capsule by placing the controller into the capsule until a yellow outline appears, and then pull the trigger. You will then be teleported to the evaluation content.
3. At first the scoring panel will be hidden. Before bringing up the panel, take a moment to look, move, and teleport or touchpad walk around your surroundings to become accustomed to the environment. Once you have an understanding of the activity within this scene, continue with the evaluation using the controls shown in Step 3. Once you feel like the ratings you have provided for all conditions best reflect your perception, press ‘Next’ to move onto the next scene. Repeat this process until you are returned to the `\_Main' scene.
4. Put down your controllers and take off the Headphones and HMD.



Touching the touchpad at the top middle moves the user in the direction that the controller is pointed.

Touching any other part of the touchpad will move the user to the right, left or backwards relative to the direction of the controller.

1. Test design statistical references:

   1. Blocking (statistics) . (2021, January 4). In Wikipedia.  
      <https://en.wikipedia.org/wiki/Blocking_(statistics)>
   2. Block design. (2021, January 4). In Wikipedia. <https://en.wikipedia.org/wiki/Block_design>
   3. Lukas Meier: Incomplete Block Designs (2021, January 4).  
      <https://stat.ethz.ch/education/semesters/as2015/anova/09_Incomplete_Blocks.pdf>
   4. Lukas Meier: Chapter 9 : Incomplete Block Designs. In ANOVA : A short Intro Using R  
      <https://stat.ethz.ch/~meier/teaching/anova/incomplete-block-designs.html>

   [↑](#footnote-ref-2)
2. See complexity\_measurement folder in AEP project in MPEG GitLab [↑](#footnote-ref-3)
3. Based on 1) DIN 1320, Akustik-Begriffe, Acoustics Terminology, 2009, Beuth Verlag, Berlin, and 2) Silzle, A. Quality taxonomies for auditory virtual environments. in 112th AES Convention 2007. Vienna, Austria [↑](#footnote-ref-4)