**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC1/SC29/WG11**

**CODING OF MOVING PICTURES AND AUDIO**

**ISO/IEC JTC1/SC29/WG11 MPEG2017/N19407**

**July 2020, Online**

|  |  |
| --- | --- |
| **Title** | **MPEG-H 3D Audio Baseline Profile Verification Test Report** |
| **Source** | **Audio Subgroup** |

# Executive summary

The MPEG-H 3D Audio Baseline Profile supports coding audio as audio channels and audio objects and is a subset of the Low Complexity Profile.

Five tests were conducted to assess the performance of the Baseline Profile of MPEG-H 3D Audio. The tests covered a range of bit rates and a range of "immersive audio" use cases (i.e. from 22.2 down to 2.0 channel presentations). Nine test sites participated in the tests with a total of 341 listeners. This resulted in a data set of 1144592 individual scores.

The statistical analysis of the test data resulted in the following conclusions:

* Test 1 measured performance for the "Ultra-HD Broadcast" use case, in which highly immersive audio material was coded at 768 kb/s and presented using 22.2 or 7.1+4H channel loudspeaker layouts. The test showed that at the bit rate of 768 kb/s, MPEG-H 3D Audio Baseline Profile easily achieves "ITU-R High-Quality Emission" quality, as needed in broadcast applications.
* Test 2 measured performance for the "HD Broadcast" or "A/V Streaming" use case, in which immersive audio material was coded at three different bit rates: 512 kb/s, 384 kb/s and 256 kb/s and presented using 7.1+4H or 5.1+2H channel loudspeaker layouts. The test showed that for all bit rates, MPEG-H 3D Audio Baseline Profile achieved a quality of "Excellent" on the MUSHRA subjective quality scale.
* Test 3 measured performance for the "High Efficiency Broadcast" use case, in which audio material was coded at three different bit rates, with specific bit rates depending on the number of channels in the material. Bitrates ranged from 256 kb/s (5.1+2H) to 48 kb/s (stereo). The test showed that for all bit rates, MPEG-H 3D Audio Baseline Profile achieved a quality of "Excellent" on the MUSHRA subjective quality scale.
* Test 4 measured performance for the "Mobile" use case, in which audio material was coded at 384 kb/s, and presented via headphones. The MPEG-H 3D Audio FD binauralization engine was used to render a virtual, immersive audio sound stage for the headphone presentation. The test showed that at 384 kb/s, MPEG-H 3D Audio Baseline Profile with binauralization achieved a quality of "Excellent" on the MUSHRA subjective quality scale.
* Test 5 measured performance for the "High Quality Immersive Music Delivery" use case in which object based immersive music is delivered to the receiver with up to 24 objects at high per object bit rates. This test used 11.1 (as 7.1+4H) as presentation format, with material coded at a rate of 1536 kb/s. The test showed that at the bit rate of 1536 kb/s, MPEG-H 3D Audio Baseline Profile easily achieves "ITU-R High-Quality Emission" quality, as needed in high quality music delivery applications.

The performance of the Low Complexity Profile of MPEG-H 3D Audio was assessed in the MPEG-H 3D Audio Verification Test Report [1]. The results for channels and objects from the Low Complexity Profile Verification Tests are re-used in this report.

Taken together, the tests provide evidence that the requirements set forth in the 3D Audio Call for Proposals [2] are fulfilled by the MPEG-H 3D Audio Baseline Profile.

**Contents**

[Executive summary 1](#_Toc44500260)

[1 Introduction 4](#_Toc44500261)

[2 Listening tests 4](#_Toc44500262)

[2.1 Test methodology 5](#_Toc44500263)

[2.2 Test material 6](#_Toc44500264)

[2.3 Test 1 "Ultra HD Broadcast" 6](#_Toc44500265)

[2.4 Test 2 "HD Broadcast" or "A/V Streaming" 7](#_Toc44500266)

[2.5 Test 3 "High Efficiency Broadcast" 8](#_Toc44500267)

[2.6 Test 4 "Mobile" 10](#_Toc44500268)

[2.7 Test 5 "High Quality Immersive Music Delivery Listening" 11](#_Toc44500269)

[3 Test plan 11](#_Toc44500270)

[3.1 Preparation of original and processed items 11](#_Toc44500271)

[3.2 Listening labs 12](#_Toc44500272)

[4 Statistical Analysis and Test Results 12](#_Toc44500273)

[4.1 Listener post-screening 12](#_Toc44500274)

[4.2 Overview 13](#_Toc44500275)

[4.3 Test 1 "Ultra HD Broadcast" 13](#_Toc44500276)

[4.4 Test 2 "HD Broadcast" or "A/V Streaming" 15](#_Toc44500277)

[4.5 Test 3 "High Efficiency Broadcast" 16](#_Toc44500278)

[4.6 Test 4 "Mobile" 20](#_Toc44500279)

[4.7 Test 5 "High Quality Immersive Music Delivery Listening" 21](#_Toc44500280)

[5 Conclusion 22](#_Toc44500281)

[6 References 23](#_Toc44500282)

[Annex A Post-screening and statistical analysis 24](#_Toc44500283)

[A.1 Post-screening analysis 24](#_Toc44500284)

[A.2 Statistical analysis 25](#_Toc44500285)

[A.3 Post-screening for Test 5 25](#_Toc44500286)

# Introduction

The MPEG-H 3D Audio Baseline Profile supports coding audio as audio channels and audio objects and is a subset of the Low Complexity Profile.

Each content type (channels, objects) can be used alone or in combination with the other. The use of audio channel and object groups allows for interactivity or personalization of a program, e.g. by selecting different language tracks or adjusting the gain or position of the objects during rendering in the MPEG-H decoder.

In MPEG-H 3D Audio the format of audio program content and the coded representation that is transmitted is independent of the consumer’s playback setup. The MPEG-H 3D Audio decoder renders the bitstream to a number of standard speaker configurations as well as for speakers that are not placed in the ideal positions. Binaural rendering of sound for headphone listening is also supported.

The standard may be used in a wide variety of applications including stereo and surround sound storage and transmission. Its support for interactivity and immersive sound is important to satisfy the requirements of next-generation media delivery, particularly new television broadcast systems and entertainment streaming services.

For example, in TV broadcasting, commentary or dialogue may be sent as audio objects and combined with an immersive channel bed in the MPEG-H 3D Audio decoder. This allows efficient transmission of dialogue in multiple languages and also allows the listener to adjust the balance between dialogue and other sound elements to his or her preference. This concept can be extended to other elements not normally present in a broadcast, such as audio description for the visually impaired, director's commentary, or to dialogue from participants in sporting events.

The MPEG-H 3D Audio specification is published as ISO/IEC 23008-3:2019, Second Edition. The requirements for the work item are shown in N16584, Annex 2 [1]. Amendment 2, specifying the Baseline Profile of MPEG-H 3D Audio is defined in N19402 [3]. An integration of the base document and all amendments, as MPEG-H 3D Audio Third Edition, is expected to be published in early 2021.

Verification tests were conducted to assess the subjective quality of the Second Edition technology including the 3D Audio Baseline profile, corrections and improvements specified in Amendment 2. Five tests were conducted to assess performance across a range of bit rates (i.e. from 1536 kb/s to 48 kb/s), a range of "immersive" use cases (i.e. from 22.2 to 2.0 channel presentations) as well as for high quality immersive music streaming. Nine test sites participated in the tests with a total of 341 listeners. This resulted in a large data set of 1144592 individual scores

# Listening tests

The five listening tests (Test 1, Test 2, Test 3, Test 4 and Test 5) were designed to assess the performance of the Baseline Profile of MPEG-H 3D Audio for five important and distinct use cases in which content is broadcast or streamed to the user.

Test 1 assesses performance for the "Ultra HD Broadcast" use case, in which it is expected that video is Ultra HD and audio is highly immersive. Considering that such video content requires considerable bit rate, it is appropriate to allocate a proportional bit rate to audio. This test used 22.2 and 11.1 (as 7.1+4H) presentation formats, with material coded at a rate of 768 kb/s.

Test 2 assesses performance for the "HD Broadcast" or "A/V Streaming" use case, in which video has HD resolution and audio is immersive: 11.1 channel (as 7.1+4H) or 7.1 (as 5.1+2H) presentation formats. To assess codec performance for interactive content, the test contained items with multiple language tracks, that were all transmitted and the choice of the rendered language track was switched at predefined times by an automation at the decoder. For streaming and even for broadcast, there is increasing demand to deliver high-quality content at lower bitrates. In order to get a sense of the rate-distortion performance of 3D Audio, this test coded audio at three intermediate bit rates: 512 kb/s, 384 kb/s and 256 kb/s.

Test 3 assesses performance for the "High Efficiency Broadcast" use case, in which content is broadcast or streamed at very low bit rates. In order to get a sense of the rate-distortion performance of 3D Audio and to address a broader range of immersive to traditional content presentation formats, this test coded audio at three intermediate bit rates, from 256 kb/s for 5.1+2H presentation format to 48 kb/s for 2.0 presentation format.

Test 4 assesses performance for the "Mobile" use case, in which content is delivered to a mobile platform such as a smartphone. Since audio playback with such platforms is typically done via headphones, this test was conducted using headphone presentation. It used the immersive content from Test 2 (i.e. 7.1+4H and 5.1+2H presentation format) but rendered for headphone presentation using the MPEG-H 3D Audio FD binauralization engine. This permits the user to perceive a fully immersive sound stage with sound sources appropriately virtualized in the 3D space.

Test 5 assesses "High Quality Immersive Music Delivery" use case in which object based immersive music is delivered to the receiver with up to 24 objects at high per object bit rates. This test used 11.1 (as 7.1+4H) presentation format, with material coded at a rate of 1536 kb/s.

Listening for Test 1, Test 2, Test 3 and Test 5 was conducted in acoustically isolated rooms using loudspeakers for presentation. A single subject was in the room during a given test session. Listening for Test 4 was conducted in acoustically isolated sound booths using headphones for presentation. A single subject was in the booth during a given test session.

## Test methodology

**BS.1116**

Test 1 and Test 5 used the Recommendation ITU-R BS.1116-3 double-blind triple-stimulus with hidden reference test methodology [4]. This methodology is appropriate for assessment of systems having small impairments, and so was only used for these tests in which the coding bitrate of 768 kb/s and 1536 kb/s would ensure that coding artefacts would be very small. The subjective response is recorded on a scale ranging from 1 to 5, with one decimal digit.

The descriptors and the score associated with each descriptor of the subjective scale are shown here:

Imperceptible (5.0)

Perceptible, but not annoying (4.0)

Slightly annoying (3.0)

Annoying (2.0)

Very annoying (1.0)

Listener instructions for the BS.1116 test are given in N16584, Annex 6 [1].

**MUSHRA**

Test 2, Test 3 and Test 4 used the Recommendation ITU-R BS.1534-3 MUSHRA test method [5]. This methodology is appropriate for assessment of systems with intermediate quality levels. The subjective response is recorded on a scale ranging from 0 to 100, with no decimal digits.

The descriptors and the range of scores associated with each descriptor of the subjective scale are shown here:

Excellent (80-100)

Good (60-80)

Fair (40-60)

Poor (20-40)

Bad (0-20)

Listener instructions for the MUSHRA test are given in N16584, Annex 6 [1].

## Test material

Test material was either channel-based or channels plus objects. The number and layout of the channel-based signals is indicated as numChannels.numLFE or as numMid.numLFE + numHigh. The latter is used where there might be some confusion between a purely mid-plane layout and a mid plus high layout, e.g. 5.1+2H, where the "numHigh" is followed by "H" to indicate the high plane. The terms used in this designation are as follows:

|  |  |
| --- | --- |
| numChannels | The total number of full-range channels, encompassing low, mid and high planes. |
| numLFE | The number of LFE channels |
| numMid | The number of mid-plane full-range channels. |
| numHigh | The number of high-plane full-range channels. |

The filenames for each test item are given in N16584, Annex 5 [1].

## Test 1 "Ultra HD Broadcast"

The following table describes the parameters for Test 1.

|  |  |
| --- | --- |
| Test Goal | Demonstrate ITU-R High-Quality Emission |
| Test Methodology | BS.1116 |
| Presentation | Loudspeaker |
| Content Formats | See Test Material, Test 1 table. |
| Content Specialties | Switch group with 3 languages that cycles through the languages (item T1\_6). |
| Reference | See Test Material, Test 1 table. |
| Test Conditions | 1. Hidden Reference 2. Full decoding of all items and rendering to presentation format. |
| Anchor | None |
| Listening Position | Sweet spot |
| Test Items | See Test Material, Test 1 table. |
| Bit Rates | 768 kb/s |
| Notes | All formats in one test  Low Complexity and Baseline Profiles |
| Requirements addressed | * High Quality * Localization and Envelopment * Audio program inputs: 22.2, discrete audio objects * Interactivity |

The following material was used in Test 1.

* For T1\_2, item was created by rendering objects ("steps") to a 22.2 channel bed.
* For T1\_5, reference was created by rendering all objects to the channel bed.
* For T1\_6, reference was created by rendering the 3 commentary objects to the channel bed such that it transitions from one language to the next.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Item** | **Content Format** | **Presentation Format** | **Item Name** | **Item Description** |
| T1\_1 | 22.2 | 22.2 | Funk | Drums, guitar, bass |
| T1\_2 | 22.2 | 22.2 | Rain with steps | Rain with steps (steps as obj) |
| T1\_3 | 22.2 | 22.2 | Swan Lake | Tchaikovsky with full orchestra |
| T1\_4 | 22.2 | 22.2 | This is SHV | Trailer for 8K Super Hi-Vision |
| T1\_5 | 7.1+4H + 3 obj | 7.1+4H | Sintel Dragon Cave (3 obj) | Fighting film scene with score |
| T1\_6 | 7.1+4H + 3 obj | 7.1+4H | DTM Car Race (3 obj, commentary languages) | Car race with 3 commentaries in 3 different languages |
| T1\_7 | 7.1+4H | 7.1+4H | Birds Paradise | Ambience with birds |
| T1\_8 | 7.1+4H | 7.1+4H | Musica Floria | String ensemble recorded in medieval church |

## Test 2 "HD Broadcast" or "A/V Streaming"

The following table describes the parameters for Test 2.

|  |  |
| --- | --- |
| Test Goal | Demonstrate MUSHRA "Excellent" (80+) |
| Test Methodology | MUSHRA |
| Presentation | Loudspeaker |
| Content Formats | See Test Material, Test 2 table. |
| Content Specialties | Switch group with 2 languages that cycles through the languages (item T2\_6). |
| Reference | See Test Material, Test 2 table. |
| Test Conditions | 1. Hidden Reference 2. 3D Audio at 512 kb/s 3. 3D Audio at 384 kb/s 4. 3D Audio at 256 kb/s 5. Anchor 1 6. Anchor 2 |
| Anchor | Anchor 1: original, LP filtered, 7.0 kHz  Anchor 2: original, LP filtered, 3.5 kHz |
| Listening Position | Sweet spot |
| Test Items | See Test Material, Test 2 table. |
| Bit Rates | Three different bit rates as shown above |
| Notes | All formats in one test  Low Complexity and Baseline Profiles |
| Requirements addressed | * High Quality * Localization and Envelopment * Audio program inputs: channel-based PCM, discrete audio objects * Interactivity |

The following material was used in Test 2.

* For T2\_1, item was created by rendering objects to a 7.1+4H channel bed.
* For T2\_2, item was created by rendering the 3 commentary objects to the channel bed such that it transitions from one language to the next.
* For T2\_5, reference was created by rendering object to 5.1+2H channel bed.
* For T2\_6, reference was created by rendering the 2 commentary objects to the channel bed such that it transitions from the English commentary to the German commentary.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Item** | **Content Format** | **Presentation Format** | **Item Name** | **Item Description** |
| T2\_1 | 7.1+4H | 7.1+4H | Sintel Dragon Cave | Fighting film scene with score |
| T2\_2 | 7.1+4H | 7.1+4H | DTM Car Race | Car race with 3 commentaries in 3 different languages |
| T2\_3 | 7.1+4H | 7.1+4H | Birds Paradise | Ambience with birds |
| T2\_4 | 7.1+4H | 7.1+4H | Musica Floria | String ensemble recorded in medieval church |
| T2\_5 | 5.1+2H + 3 obj | 5.1+2H | Sintel Dragon Cave | Fighting film scene with score |
| T2\_6 | 5.1+2H + 2 obj | 5.1+2H | Handball Commentary | Sports with commentaries in 2 different languages |
| T2\_7 | 5.1+2H | 5.1+2H | Blug Hendrix Beat | Live rock concert |
| T2\_8 | 5.1+2H | 5.1+2H | Song World Percussion | Pop Music with drums |

## Test 3 "High Efficiency Broadcast"

The following table describes the parameters for Test 3.

|  |  |
| --- | --- |
| Test Goal | Demonstrate MUSHRA "Good" quality at low bit rates |
| Test Methodology | MUSHRA |
| Presentation | Loudspeaker |
| Content Formats | See Test Material, Test 3 table. |
| Content Specialties | None |
| Reference | See section on Test 3 Material, above. |
| Test Conditions | 1 Hidden Reference   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | |  |  | 5.1+2H | 5.1 | 2.0 |  | | 2 | 3D Audio | 256 kb/s | 180 kb/s | 80 kb/s |  | | 3 | 3D Audio | 192 kb/s | 144 kb/s | 64 kb/s |  | | 4 | 3D Audio | 144 kb/s | 128 kb/s | 48 kb/s |  |   5 Anchor 1  6 Anchor 2 |
| Anchor | Anchor 1: original, LP filtered, 7.0 kHz  Anchor 2: original, LP filtered, 3.5 kHz |
| Listening Position | Sweet spot |
| Test Items | See Test Material, Test 3 table. |
| Bit Rates | As in Test Conditions row of this table. |
| Notes | All formats in one test  Low Complexity and Baseline Profiles  No interactivity  No dynamic objects |
| Requirements addressed | * High Quality * Localization and Envelopment * Audio program inputs: channel-based PCM, discrete audio objects |

The following material was used in Test 3.

* For T3\_1 and T3\_2, item was created by rendering all objects to a 5.1+2H channel bed.
* For T3\_2, only English commentary was used.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Item** | **Content Format** | **Presentation Format** | **Item Name** | **Item Description** |
| T3\_1 | 5.1+2H | 5.1+2H | Sintel Dragon Cave | Fighting film scene with score |
| T3\_2 | 5.1+2H | 5.1+2H | Handball Commentary | Sports with commentary |
| T3\_3 | 5.1+2H | 5.1+2H | Blug Hendrix Beat | Live rock concert |
| T3\_4 | 5.1 | 5.1 | Mancini | Movie score with brass |
| T3\_5 | 5.1 | 5.1 | Bach 565 | Bach Toccata d minor |
| T3\_6 | 5.1 | 5.1 | Sedambonjou Salsa | Latin music with brass and percussions |
| T3\_7 | 2.0 | 2.0 | Susanne Vega (te8) | Suzanne Vega, Tom’s Diner |
| T3\_8 | 2.0 | 2.0 | Tracy Chapman (te9) | Tracy Chapman |
| T3\_9 | 2.0 | 2.0 | Hockey | Hockey Game |

Note: Items T3\_5, T3\_6 and T3\_9 were kindly provided by EBU.

## Test 4 "Mobile"

The following table describes the parameters for Test 4.

|  |  |
| --- | --- |
| Test Goal | Demonstrate MUSHRA "excellent" (80+) |
| Test Methodology | MUSHRA |
| Presentation | Headphones |
| Content Formats | Same as in Test 2, "HD Broadcast" or "A/V Streaming" |
| Content Specialties | None |
| Reference | Channels:  PCM original item processed by BRIR as full convolution.  Objects: If items contain objects, the objects are rendered to Presentation Format and then processed by BRIR as full convolution.  BRIR are the same BRIR as was used in MPEG-H 3D Audio CfP |
| Test Conditions | 1. Hidden Reference 2. C/O: MPEG-H using FD binauralization engine 3. Anchor 1 4. Anchor 2 |
| Anchor | Anchor 1: Anchor 1 from Test 2, then processed by BRIR  Anchor 2: Anchor 2 from Test 2, then processed by BRIR |
| Listening Position | N/A |
| Test Items / Bit Rates | Use 384 kb/s bitstreams from Test 2 |
| Restrictions | None |
| Notes | All formats in one test |
| Requirements addressed | * High Quality * Localization and Envelopment * Audio program inputs: channel-based PCM, discrete audio objects * Rendering for Headphone Listening * HRTF Personalization |

Test 4 used the same material as Test 2. More specifically, in Test 4 the 3D Audio decoder processed the Test 2 bitstreams to create a binauralized stereo result. The binauralization used a Binaural Room Impulse Response (BRIR), specifically, the same BRIR as was used in the MPEG‑H 3D Audio Call for Proposals [2]. This BRIR was recorded in the Mozart listening room at Fraunhofer IIS.

## Test 5 "High Quality Immersive Music Delivery Listening"

|  |  |
| --- | --- |
| **Test Goal** | **Demonstrate ITU-R High-Quality Emission** |
| Test Methodology | BS.1116 |
| Presentation | Loudspeaker |
| Content Formats | See Test Material, Test 5 table. |
| Content Specialties | Object based immersive music. |
| Reference | See Test Material, Test 5 table. |
| Test Conditions | 1. Hidden Reference 2. Full decoding of all items and rendering to presentation format. |
| Anchor | None |
| Listening Position | Sweet spot |
| Test Items | See Test Material, Test 5 table. |
| Bit Rate | 1536 kb/s |
| Notes | Baseline Profile |
| Requirements addressed | * High Quality Immersive Music Delivery * Localization and Envelopment * Audio program inputs: 24 discrete audio objects |

The following material was used in T5:

* All references were created by rendering all objects to the 7.1+4H speaker channels.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Item** | **Content Format** | **Presentation Format** | **Item Name** | **Item Description** |
| T5\_1 | 24 Objects | 7.1+4H | Then\_Funk\_01 | Drums, guitar, bass |
| T5\_2 | 24 Objects | 7.1+4H | Then\_Funk\_02 |
| T5\_3 | 24 Objects | 7.1+4H | Then\_Funk\_03 |
| T5\_4 | 24 Objects | 7.1+4H | SpaceInvaders\_01 | Electronic Music |
| T5\_5 | 24 Objects | 7.1+4H | SpaceInvaders\_02 |
| T5\_6 | 24 Objects | 7.1+4H | SpaceInvaders\_03 |
| T5\_7 | 24 Objects | 7.1+4H | SpaceInvaders\_04 |

# Test plan

## Preparation of original and processed items

Original items were provided by ARL, EBU, ETRI, Fraunhofer IIS, FTV, NHK, Orange and Qualcomm. They were limited to not more than 20 seconds duration and were edited to have "fade-in" and "fade-out" at beginning and end.

All channel and channel plus object test items were processed, i.e. encoded/decoded and low-pass filtered, by Fraunhofer IIS.

## Listening labs

The following table shows the listening labs that participated in each listening test. The number of subjects participating from each lab in a given test is shown in the table entries; a blank entry indicates no participation. The total number of listeners in each test is shown in the last row of the table.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Test Lab** | **Test 1** | **Test 2** | **Test 3** | **Test 4** | **Test 5** |
| ETRI |  | 12 |  | 12 |  |
| FhG-IIS | 24 | 24 | 29 | 28 | 18 |
| NHK | 18 | 18 | 18 |  |  |
| Nokia |  |  | 10 | 12 |  |
| Orange |  |  |  | 9 |  |
| Qualcomm | 16 | 15 | 16 | 16 |  |
| Sony | 11 |  |  |  | 14 |
| Gaudio |  |  |  |  | 13 |
| Dolby |  |  |  |  | 8 |
| **Total** | **69** | **69** | **73** | **77** | **53** |

For Test 1, Test 2, Test 3 and Test 5, the listening labs all had high-quality listening rooms that were calibrated to conform to the criteria set forth in BS.1116 and also calibrated to be perceptually similar to each other. Hence, the test lab subjective results can be pooled together for each of the tests. The loudspeaker positions used when presenting the various test item is shown in N16584, Annex 5, Table 2 [1], specifically the loudspeaker azimuth (A+000) and elevation (E+00) angles are shown under the heading "Label."

For Test 4, the listening labs used acoustically isolating sound booths and high-quality headphones.

# Statistical Analysis and Test Results

## Listener post-screening

**Test 1**

Test 1 used the BS.1116 test methodology [4]. For each listener in Test 1, post-screening of listener responses was based on the listener’s ability to correctly differentiate the Hidden Reference from the System under Test, which is the procedure recommended in BS.1116-3. The exact procedure used is described in Annex A.

The post-screening procedures computes the statistic which is the 95% point of the cumulative distribution of the listener Diff Grades, which are assumed to have the Student t distribution. If for the listener *i*, then we conclude, with a 95% level of significance, that the listener cannot reliably differentiate between the Hidden Reference and the System under Test, and listener responses for the 12 test items are removed from consideration.

**Test 2, Test 3, Test 4**

Test 2, Test 3 and Test 4 used the MUSHRA test methodology [5]. For each listener in each test, post-screening of listener responses was based on scores for Hidden Reference and Low Pass filtered anchors. The procedure is as follows:

If, for any test item in a given test, either of the following criterion are *not* satisfied:

* The listener score for the Hidden Reference is greater than or equal to 90 (i.e. HR >= 90)
* The listener score for the Hidden Reference, the 7.0 kHz lowpass anchor and the 3.5 kHz lowpass anchor are monotonically decreasing (i.e. HR >= LP70 >= LP35).

then all listener responses in that test are removed from consideration.

**Test 5**

Test 5 used the BS.1116 test methodology [4]. However, this configuration of Baseline profile is nearly perceptually transparent, and applying strict post-screening rules would result in excluding many of the expert listeners. Therefore, it was decided not to apply post-screening. In either case, the mean score of the system under test does not significantly change and the system under test qualifies as "ITU-R High-Quality Emission", as described in Annex A.3.

## Overview

Statistical analysis was performed on subjective scores remaining after listener post-screening. Details of the statistical analysis are given in Annex A. After applying post-screening there were at least 35 listeners for Test 1, Test 2 Test 3 and Test 4. These numbers far exceed the BS.1116-3 and BS.1534-3 recommendations of at least 20 listeners per test. For Test 5 no post-screening was applied, and scores from all 45 participating listeners were included in the statistical analysis.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Test** | **Test 1** | **Test 2** | **Test 3** | **Test 4** | **Test 5** |
| **Subjective Scores in**  **statistical analysis** | **35** | **43** | **44** | **68** | **53** |

For Test 1 and Test 5, a Diff Grade was computed (as Hidden Reference – System under Test scores) and statistics were computed on the Diff Grade. In addition, statistical analysis was performed on absolute scores for Hidden Reference and the System under Test. For Test 2, Test 3 and Test 4, statistics were computed on the absolute MUSHRA scores.

The tables in this section show, for each System under Test (Sys), the mean score (Mean) as averaged over all listeners (after post-screening) and all test items. For each result, the 95% confidence interval on the mean score was computed, and the table shows the upper (High) and lower (Low) limits of the 95% confidence interval.

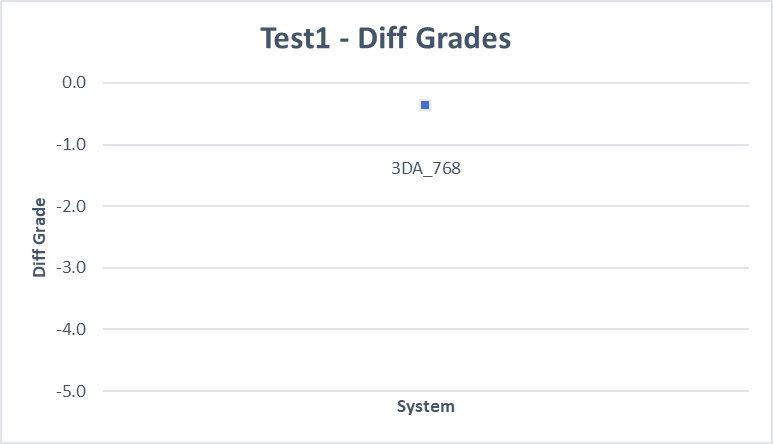
Note that the 95% confidence interval is shown in every plot, but when retaining the full subjective scale, the interval is obscured by the mark used to indicate the mean value. However, 95% confidence intervals are shown in the tabular presentation of scores.

## Test 1 "Ultra HD Broadcast"

The following table shows the mean score for 3D Audio system operating at 768 kb/s (3DA\_768) and the associated high and low 95% confidence interval limits on the mean.

|  |  |  |  |
| --- | --- | --- | --- |
| Sys | High | Low | Mean |
| 3DA\_768 | -0.33 | -0.42 | -0.37 |

The following is a plot of the mean score and 95% confidence interval. The confidence interval is plotted, but is so small that it is within the size of the marker used for the mean.

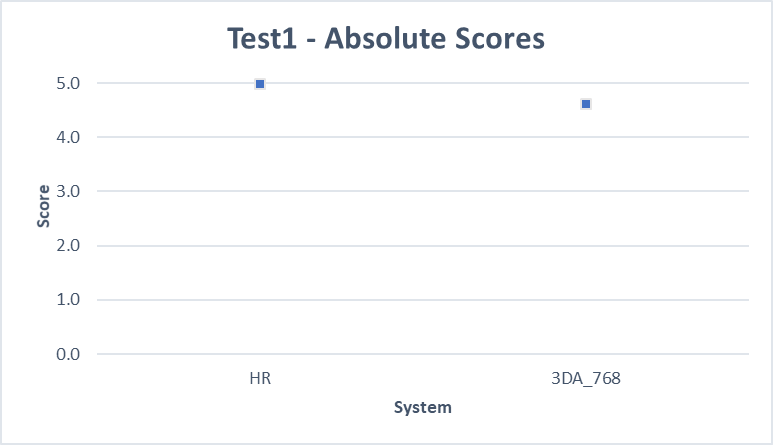


The following table and plot show the mean score for 3D Audio system operating at 768 kb/s (3DA\_768), the Hidden Reference (HR) and the associated high and low 95% confidence interval limits on the mean for each condition.

For the 3DA\_768, the absolute score is not lower that 4.6 at the 95% level of confidence, which is well above the 4.0 limit recommended in Recommendation ITU-R BS.1548-7 [6] for "High-quality emission" for broadcast applications (indicated by red line in the plot). Recommendation ITU-R BS.1548-7, Annex 2 Section 2.1.1.1 "High-quality emission" states "Ideally, the quality of the sound reproduced after decoding will be subjectively similar to the original signal for most types of audio programme material. Using the triple stimuli double blind with hidden reference test, described in Recommendation ITU-R BS.1116, this requires mean values consistently higher than 4 on the Recommendation ITU-R BS.1116 5-grade impairment scale at the reference listening position."

|  |  |  |  |
| --- | --- | --- | --- |
| Sys | High | Low | Mean |
| HR | 4.99 | 4.97 | 4.98 |
| 3DA\_768 | 4.65 | 4.56 | 4.61 |

The following is a plot of the mean scores and 95% confidence intervals. The confidence intervals are plotted, but are so small that they are within the size of the marker used for the mean. The red line shows the ITU-R requirement for "high-quality emission," i.e. mean value of 4.0.

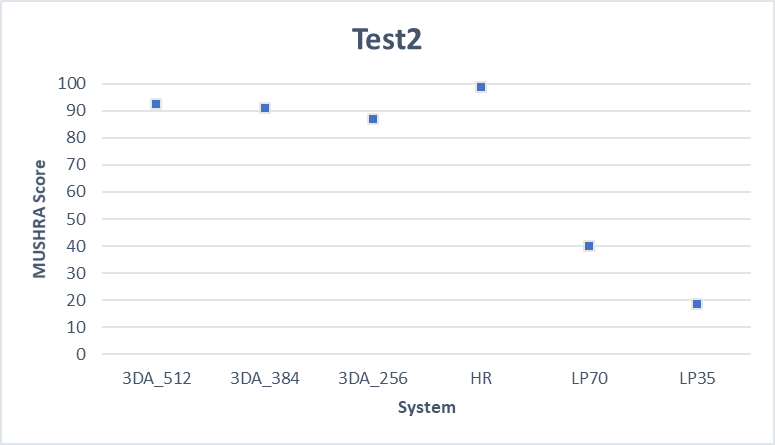


## Test 2 "HD Broadcast" or "A/V Streaming"

The following table shows the mean score for 3D Audio operating at 512 kb/s (3DA\_512), 384 kb/s (3DA\_384), 256 kb/s (3DA\_256), the Hidden Reference (HR), the 7.0 kHz low pass anchor (LP70) and 3.5 kHz low pass anchor (LP35), and the associated high and low 95% confidence interval limits on the mean for each condition.

|  |  |  |  |
| --- | --- | --- | --- |
| Sys | High | Low | Mean |
| 3DA\_512 | 93.36 | 91.74 | 92.55 |
| 3DA\_384 | 91.88 | 90.27 | 91.08 |
| 3DA\_256 | 88.08 | 85.79 | 86.93 |
| HR | 99.18 | 98.66 | 98.92 |
| LP70 | 41.35 | 39.02 | 40.18 |
| LP35 | 19.40 | 17.83 | 18.62 |

The following is a plot of the mean scores and 95% confidence intervals. The confidence intervals are plotted, but are so small that they are within the size of the marker used for the mean.

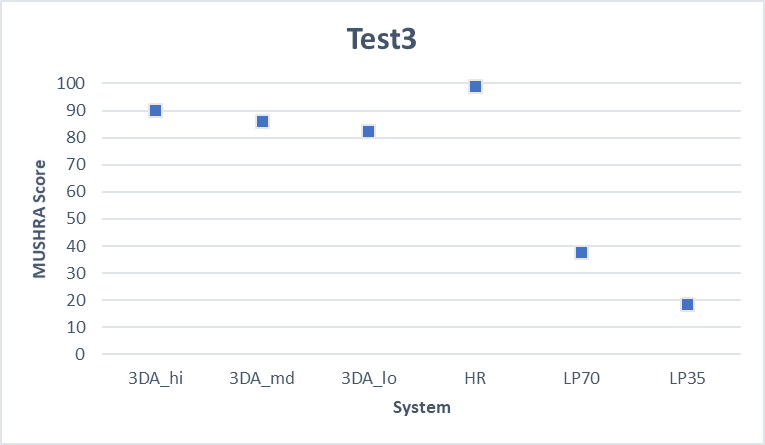
****

## Test 3 "High Efficiency Broadcast"

The following table shows the mean score for 3D Audio operating at three bit rates: 3DA\_hi, 3DA\_mid, 3DA\_lo, the Hidden Reference (HR), the 7.0 kHz low pass anchor (LP70) and 3.5 kHz low pass anchor (LP35), and the associated high and low 95% confidence interval limits on the mean for each condition. The specific bit rates for each test item for each of the three rates (hi, mid, lo) are given in the table in Section 2.5.

|  |  |  |  |
| --- | --- | --- | --- |
| Sys | High | Low | Mean |
| 3DA\_hi | 91.49 | 89.25 | 90.37 |
| 3DA\_md | 87.64 | 84.99 | 86.31 |
| 3DA\_lo | 83.94 | 80.82 | 82.38 |
| HR | 99.39 | 98.95 | 99.17 |
| LP70 | 39.08 | 36.21 | 37.65 |
| LP35 | 19.66 | 17.54 | 18.60 |

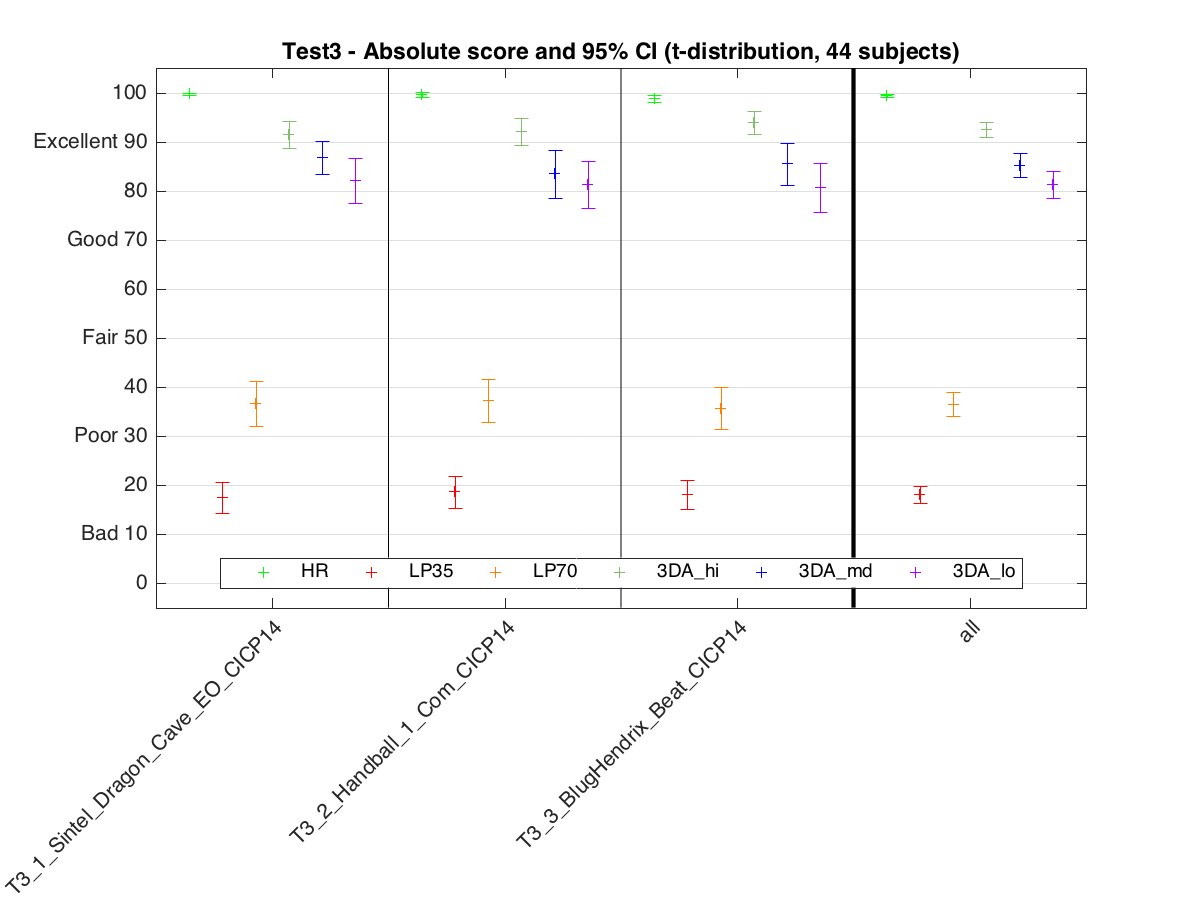
The following is a plot of the mean scores and 95% confidence intervals. The confidence intervals are plotted, but are so small that they are within the size of the marker used for the mean.



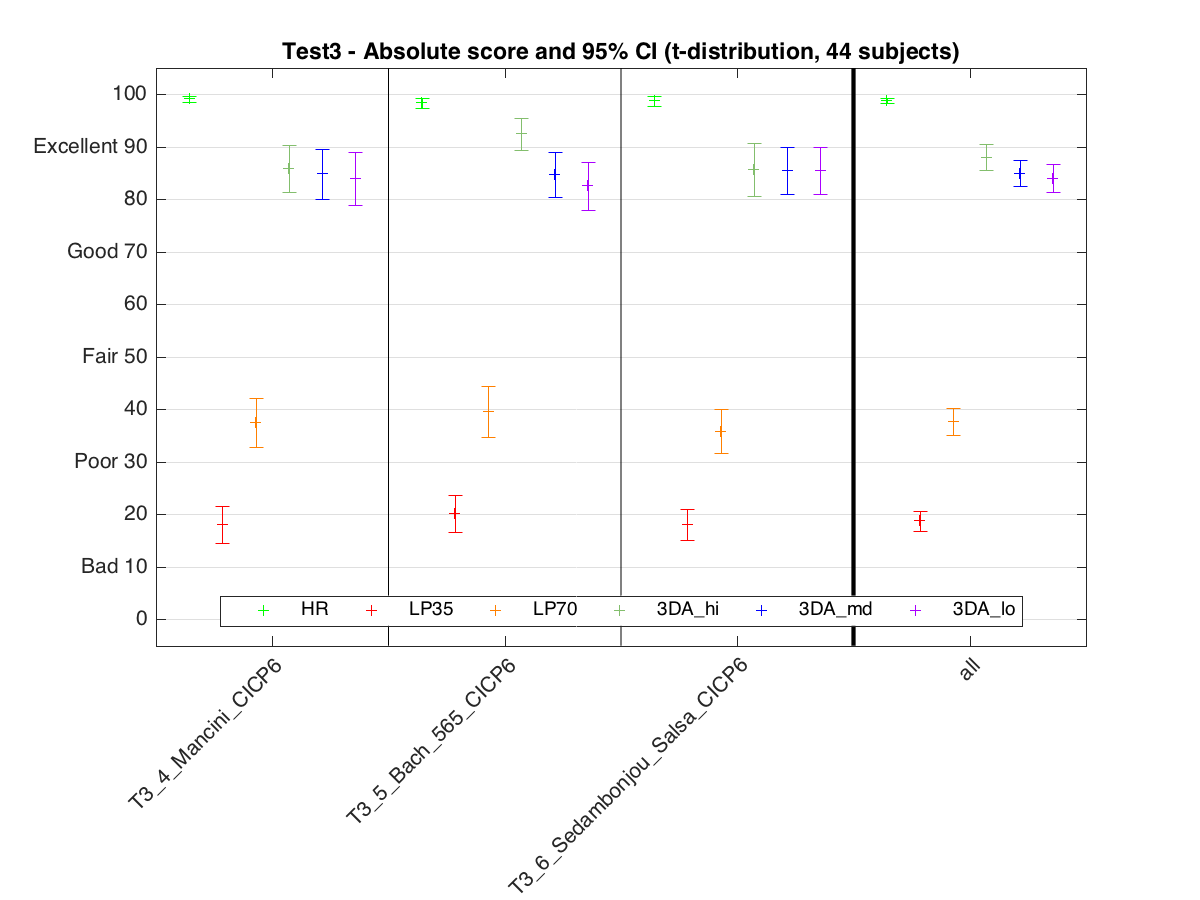
Since this test used a range of content formats for the test items and coded each format with a range of bit rates, the following table and plots present the performance of 3D Audio for each content format for the three (hi, mid, lo) coding bit rates.

|  |  |  |  |
| --- | --- | --- | --- |
| Content | High Rate | Mid Rate | Low Rate |
| Stereo | 90.60 ± 1.68 | 88.68 ± 1.98 | 81.83 ± 2.81 |
| 5.1 | 88.00 ± 2.47 | 85.02 ± 2.52 | 84.02 ± 2.63 |
| 5.1+2 | 92.50 ± 1.50 | 85.23 ± 2.36 | 81.29 ± 2.71 |

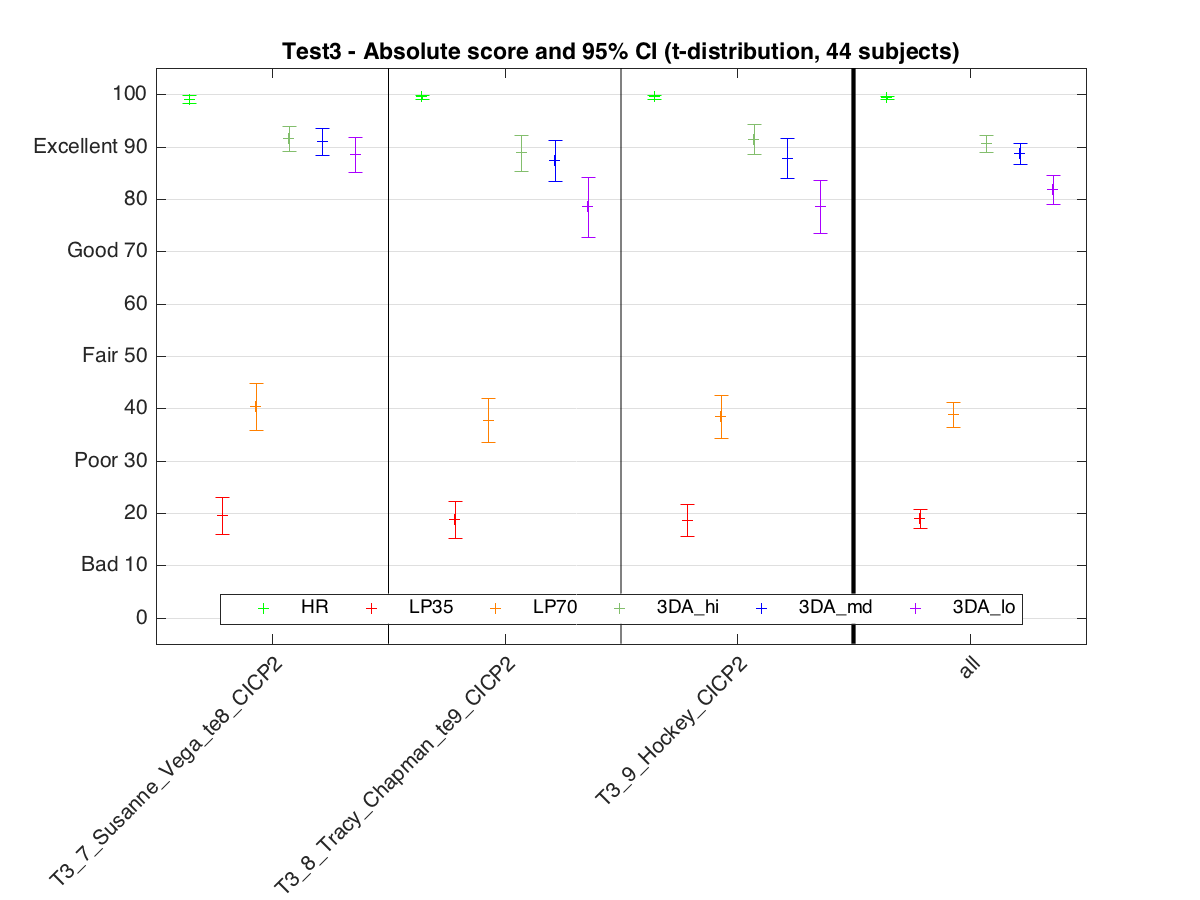
The following plot shows the performance for 5.1+2H layout (CICP 14) immersive content.



The following plot shows the performance for 5.1 layout (CICP 6) content.



The following plot shows the performance for stereo (CICP 2) content.

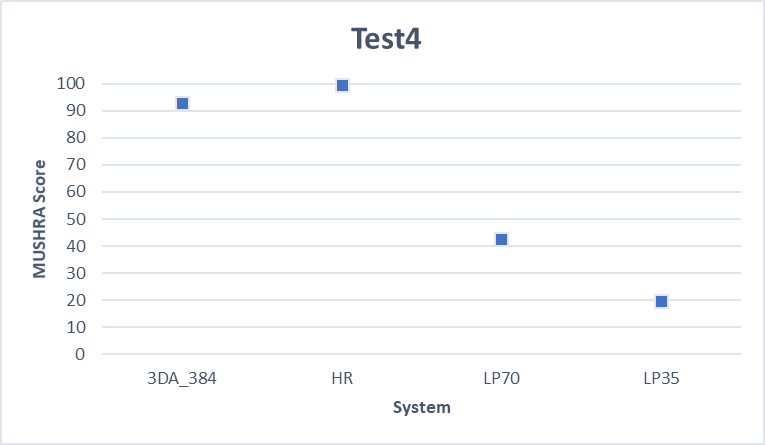


## Test 4 "Mobile"

The following table shows the mean score for 3D Audio operating at 384 kb/s (3DA\_384), the Hidden Reference (HR), the 7.0 kHz low pass anchor (LP70) and 3.5 kHz low pass anchor (LP35), and the associated high and low 95% confidence interval limits on the mean for each condition.

|  |  |  |  |
| --- | --- | --- | --- |
| Sys | High | Low | Mean |
| 3DA\_384 | 93.40 | 92.14 | 92.77 |
| HR | 99.49 | 99.16 | 99.32 |
| LP70 | 43.69 | 41.46 | 42.58 |
| LP35 | 20.67 | 18.97 | 19.82 |

The following is a plot of the mean scores and 95% confidence intervals. The confidence intervals are plotted, but are so small that they are within the size of the marker used for the mean.

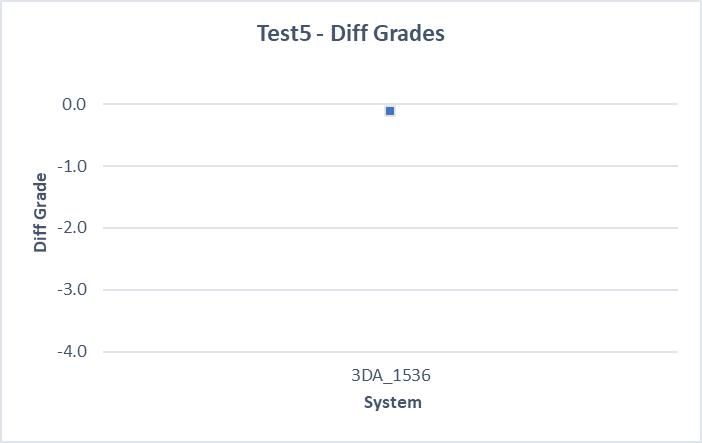


## Test 5 "High Quality Immersive Music Delivery Listening"

The following table shows the mean score for Baseline Profile of the 3D Audio system operating at 1536 kb/s (mpegh) and the associated high and low 95% confidence interval limits on the mean.

|  |  |  |  |
| --- | --- | --- | --- |
| Sys | High | Low | Mean |
| 3DA\_1536 | -0.09 | -0.15 | -0.12 |

The following plot shows the per-item and overall mean scores and 95% confidence intervals.

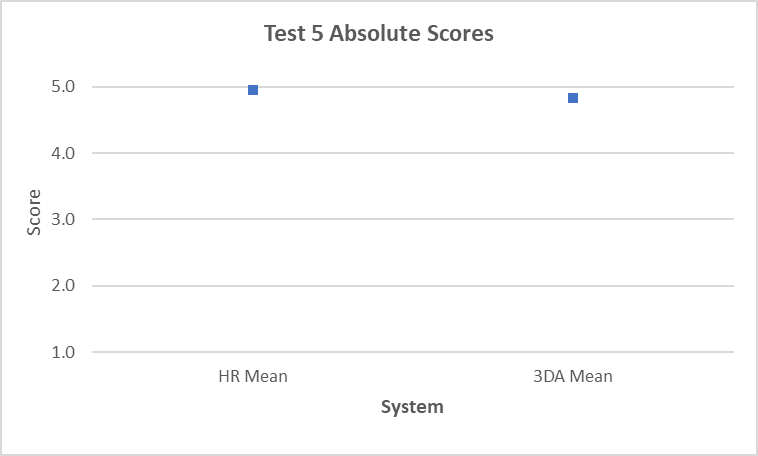


The following table and plot show the mean score for the Baseline Profile of the 3D Audio system operating at 1536 kb/s (orange), the Hidden Reference (blue) and the associated high and low 95% confidence interval limits on the mean for each condition.

|  |  |  |  |
| --- | --- | --- | --- |
| Sys | High | Low | Mean |
| HR | 4.97 | 4.92 | 4.95 |
| 3DA\_1536 | 4.85 | 4.80 | 4.82 |

For 3DA\_1536, the absolute score is not lower than 4.8 at the 95% level of confidence, which is well above the 4.0 limit recommended in ITU-R BS.1548-4 for "High-quality emission" for broadcast applications.

The following is a plot of the mean scores and 95% confidence intervals. The ITU-R requirement for "high-quality emission" is a mean score consistently above 4.0.



# Conclusion

This report provides details on four tests that were conducted to assess the performance of the Baseline Profile of MPEG-H 3D Audio. The tests covered a range of bit rates and a range of "immersive audio" use cases (i.e. from 22.2 down to 2.0 channel presentations).

The statistical analysis of the test data resulted in the following conclusions:

* Test 1 measured performance for the "Ultra-HD Broadcast" use case, in which highly immersive audio material was coded at 768 kb/s and presented using 22.2 or 7.1+4H channel loudspeaker layouts. The test showed that at the bit rate of 768 kb/s, Baseline Profile of MPEG-H 3D Audio Baseline Profile easily achieves "ITU-R High-Quality Emission" quality, as needed in broadcast applications.
* Test 2 measured performance for the "HD Broadcast" or "A/V Streaming" use case, in which immersive audio material was coded at three different bit rates: 512 kb/s, 384 kb/s and 256 kb/s and presented using 7.1+4H or 5.1+2H channel loudspeaker layouts. The test showed that for all bit rates, Baseline Profile of MPEG-H 3D Audio Baseline Profile achieved a quality of "Excellent" on the MUSHRA subjective quality scale.
* Test 3 measured performance for the "High Efficiency Broadcast" use case, in which audio material was coded at three different bit rates, with specific bit rates depending on the number of channels in the material. Bitrates ranged from 256 kb/s (5.1+2H) to 48 kb/s (stereo). The test showed that for all bit rates, Baseline Profile of MPEG-H 3D Audio Baseline Profile achieved a quality of "Excellent" on the MUSHRA subjective quality scale.
* Test 4 measured performance for the "Mobile" use case, in which audio material was coded at 384 kb/s, and presented via headphones. The MPEG-H 3D Audio FD binauralization engine was used to render a virtual, immersive audio sound stage for the headphone presentation. The test showed that at 384 kb/s, Baseline Profile of MPEG-H 3D Audio Baseline Profile with binauralization achieved a quality of "Excellent" on the MUSHRA subjective quality scale.
* Test 5 measured performance for the "High Quality Immersive Music Delivery" use case in which object based immersive music is delivered to the receiver with up to 24 objects at high per object bit rates. This test used 11.1 (as 7.1+4H) presentation format, with material coded at a rate of 1536 kb/s. The test showed that at the bit rate of 1536 kb/s, Baseline Profile of MPEG-H 3D Audio Baseline Profile easily achieves "ITU-R High-Quality Emission" quality, as needed in high quality music delivery applications.

The performance of the Low Complexity Profile of MPEG-H 3D Audio was assessed in the MPEG-H 3D Audio Verification Test Report [1]. The results for channels and objects from the Low Complexity Profile Verification Tests are re-used in this report.

Taken together, the tests provide evidence that the requirements set forth in the 3D Audio Call for Proposals [1] are fulfilled by the MPEG-H 3D Audio Baseline Profile.

# References

1. N16584, MPEG-H 3D Audio Verification Test Report
2. N13411, "Call for Proposals for 3D Audio." Available at <http://mpeg.chiariglione.org/standards/mpeg-h/3d-audio>
3. N19402, Text of ISO/IEC 23008-3:2019 FDAM 2, 3D Audio Baseline profile, Corrections and Improvements
4. ITU-R Recommendation BS.1116-3 (02/2015), "Methods for the subjective assessment of small impairments in audio systems."
5. ITU-R Recommendation BS.1534-3 (10/2015), "Method for the subjective assessment of intermediate quality level of coding systems," also known as "MUlti Stimulus test with Hidden Reference and Anchor (MUSHRA)."
6. ITU-R Recommendation BS.1548-7 (10/2019), " User requirements for audio coding systems for digital broadcasting"

# Annex A Post-screening and statistical analysis

1. Post-screening analysis

A post-screening procedure was applied to listener data in all tests to assess the subjects’ reliability.

**BS.1116**

Test 1 used the BS.1116 test methodology. For each listener in the test, post-screening was based on the listener’s ability to correctly differentiate between the Hidden Reference and the System under Test, which is the procedure recommended in BS.1116-3.

The first step is to calculate Diff Grades (d) for each listener trial

where

is Diff Grade

is the score for the System under Test

is the score for the Hidden Reference

for

subject *i* and test item *j*.

Note that if the listener ability to correctly differentiate between the Hidden Reference and the System under Test, the listener’s Diff Grades are typically less than zero since the listener should score the Hidden Reference to 5.0 and the System under Test to less than 5.0.

A single-sided test, in which the Diff Grade has the Student t distribution, is used to assess the ability of a given listener to correctly differentiate between Hidden Reference and the System under Test. We compute the statistic :

where

is the inverse Student t distribution value, that is the point in the Student t distribution for which *α* probability is in the tails. We set *α* to10% since we which to implement single-sided t-test with a 95% level of significance (i.e. 5% in one tail).

*n* is the number of scores (i.e. 12)

is the sample standard deviation of the listener’s 12 Diff Grade scores

is the sample mean of the listener’s 12 Diff Grade scores

If the statistic for the listener *i*, then we conclude, with a 95% confidence, that the listener cannot reliably differentiate between the Hidden Reference and the System under Test, and the 12 listener responses are removed from consideration.

**MUSHRA**

Test 2, Test 3 and Test 4 use the MUSHRA test methodology. For each listener in each test, post-screening was based on listener scores for Hidden Reference and Low Pass filtered anchors. The procedure is as follows:

If, for any test item in a given test, either of the following criterion are not satisfied:

* The listener score for the hidden reference is greater than or equal to 90. That is  
   HR >= 90.
* The listener scores the hidden reference, the 7.0 kHz lowpass anchor and the 3.5 kHz lowpass anchor are monotonically decreasing. That is,  
   HR >= LP70 >= LP35.

Then all listener responses in that test are removed from consideration.

1. Statistical analysis

The statistical analysis of test scores follows standard statistical procedures. The calculation of the averages over the post-screened listener scores results in the Mean Subjective Score (MSS). The first analysis step of the results considers the calculation of the mean score , for each of the presentations:

where:

is the score of subject *i* for a given test condition *j* and test item *k.*

*N* is the number of subjects

Confidence intervals were derived from the standard deviation and the size of each sample. The 95% confidence interval for a given test condition *j* and test item *k* is given by:

where

and the sample standard deviation is given by:

With a probability of 95%, the absolute value of the difference between the experimental or sample mean score and the “true” mean score (for a very high number of observers) is within the 95% confidence interval, on condition that the distribution of the individual scores are approximately Gaussian.

Similarly, a 95% confidence interval could be calculated for each test condition. In this case, sample means and sample standard deviations are calculated over all listeners and all test items.

1. Post-screening for Test 5

Test 5 used the BS.1116 test methodology [4]. However, this configuration of Baseline profile is nearly perceptually transparent, and applying strict post-screening rules as described in Annex A.1 would result in excluding many of the expert listeners. Table A. 2 shows the overall mean scores and Table A.2 the overall mean diff score for the Baseline Profile of the 3D Audio system operating at 1536 kb/s (mpegh) and the associated high and low 95% confidence interval limits on the mean using post-screening and without post-screening.

**Table A. 1 - Overall mean scores for Test 5**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Num Listeners | Sys | High | Low | Mean |
| Without  post-screening | 53 | HR | 4.97 | 4.92 | 4.95 |
| 3DA\_1536 | 4.85 | 4.80 | 4.82 |
| With  post-screening | 15 | HR | 5.05 | 4.94 | 4.99 |
| 3DA\_1536 | 4.75 | 4.65 | 4.70 |

**Table A. 2 - Overall mean diff scores for Test 5**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sys | Post-screening | Num Listeners | High | Low | Mean |
| 3DA\_1536 | Without post-screening | 53 | -0.09 | -0.15 | -0.12 |
| 3DA\_1536 | With post-screening | 15 | -0.24 | -0.35 | -0.29 |

The absolute mean score of the system under test does not change much (<=0.13 MOS) if post‑screening is applied and the system under test qualifies as "ITU-R High-Quality Emission” for both cases.

Therefore, for Test 5 no post-screening was applied, and scores from all 53 participating listeners were included in the statistical analysis.