EBU – TECH 3309

Evaluations of Cascaded Audio Codecs

Source: B/AIM (Audio in Multimedia)

Status: EBU Report

Geneva
June 2005

Page intentionally left blank. This document is paginated for recto-verso printing

Contents

1. Introduction 5
2. Coded Combinations in a Typical Broadcast Chain 5
3. Codecs Used 6
4. Test sequences 6
5. Overview of the test methodology 7
5.1. Stage 1: Initial combination selection 7
5.2. Stage 2: Objective tests 7
5.3. Stage 3: Subjective tests 7
6. Objective tests (PEAQ) 7
7. Cascades for Subjective Tests 9
8. Sharing of Subjective Tests Workload 9
9. Subjective Tests Methodology 10
9.1. Hidden reference and hidden anchors 10
9.2. Preparation 10
9.3. Headphones 10
10. Statistical Analysis 11
10.1. Rejection of subjects 11
10.2. Number of subjects 11
11. Lessons Learned from the Test Process 11
12. Evaluation Results 12
12.1. Cascade performance as a function of input, contribution, and emission codecs 13
12.1.1. Results grouped according to input codec 14
12.1.2. Results grouped according to contribution codec 14
12.1.3. Results grouped according to emission codec 14
12.2. Overall graph of 36 cascades 15
12.3. Comparing laboratories 15
12.4. Comparing the performance of each test item over all cascades 16
12.5. Correlation of objective and subjective scores. 17
13. Conclusions 17
13.1. Caveats 18
14. APPENDIX: Graphical Results 19

Page intentionally left blank. This document is paginated for recto-verso printing

Evaluations of Cascaded Audio Codecs

EBU Committee	*First Issued*	*Revised*	*Re-issued*
BMC	2005		

Keywords: Broadcast chain, concatenation, audio codec, bit rate compression

1. Introduction

EBU members have been experiencing significant problems arising from the fact that a typical broadcasting chain may comprise a number of different audio compression and decompression schemes (codecs). This experience shows that cascading different codecs generally results in a significant overall degradation of audio quality for the end user.

The purpose of this document is to -

a) analyse typical broadcasting chains,

b) identify the most critical cases,

c) evaluate possible quality degradations and,

d) propose some guidance to the Members as to how it would be possible to avoid excessive quality degradation in practical operations.

These tests have been prompted by the fact that the last time similar tests were conducted was back in 1993 by the ITU. Since then, many new codecs have been introduced to the market and consequently new issues have arisen in typical broadcasting chains.

2. Coded Combinations in a Typical Broadcast Chain

Typically, a broadcast chain consists of the following elements:

- Source
- Contribution circuit
- Broadcast studio installation
- Secondary distribution
- Emission

At each chain element several different audio codecs can be used. For example, in the contribution element as many as 13 different codecs were identified. Project Group B/AIM reviewed some broadcast chains used at the BBC, Radio France, Polish TV, ARD and other EBU member organisations.

Taking into account all possible combinations, one would arrive at a total number of combinations over 50,000. Such a huge number is of course impracticable to analyse; therefore an attempt has been made to reduce this to a tractable number.

In the reduction process we made the following assumptions and decisions:

- A chain consists of an input format and 4 codecs.
- Only digital input formats (no analogue sources).
- Sampling rate is 48 kHz (no low sampling rates)[1].
- Only currently used low bitrate audio codecs (as considered by the Members).
- Only the broadcast chain is considered, delivery over Internet is omitted.
- Everything is in stereo.

3. Codecs Used

For each stage of broadcast chain the most typically used codecs were chosen, they are described in Table 1.

Note that the labels of the emission codecs have a –12 suffix, indicating that the full bitrate is not used for audio, as 12 kbit/s is typically allocated for extra data in the ancillary bits in a real-life broadcast multiplex. The abbreviation 's' means stereo, and 'js' means joint stereo.

TABLE 1: Description of codecs used in the tests

Label	Codec	Bit-rate & stereo mode	Type Number
L2_256s	MPEG Layer 2	256kbit/s stereo	David DIGAS & Digigram PCX-9
L2_384s	MPEG Layer 2	384kbit/s stereo	David DIGAS & Digigram PCX-9
MP3_128js	MPEG Layer 3	128kbit/s joint-stereo	Lame Encoder Software v3.91
L3_128js	MPEG Layer 3	128kbit/s joint-stereo	MAYAH Centauri 3001
MiniDisc	Sony ATRAC	384kbit/s stereo	ATRAC4 Disc recorder
WM_128s	Microsoft Windows Media	128kbit/s stereo	9
ADPCM_256s	Audio Processing technology	256kbits/s stereo	AETA Codec 'MicDA 4SB'-analogue interfaces
AAC_128s	Advanced Audio Coding	128 kbit/s stereo	MAYAH Centauri 3001
L2_256s-12	MPEG Layer 2	256 kbit/s stereo (12 kbit/s ancillary data)	AVT 'Magic ISDN' Codec
L2_192js-12	MPEG Layer 2	192 kbit/s joint-stereo (12 kbit/s anc. Data)	AVT 'Magic ISDN' Codec
L2_128js-12	MPEG Layer 2	128 kbit/s joint-stereo (12 kbit/s anc. Data)	AVT 'Magic ISDN' Codec

The codecs are later referred to by a letter. The key to this, and the stage at which each codec is used, is shown in Table 2.

TABLE 2: Key to codecs used at each broadcast stage.

Input format		Contribution		Studio		Distribution		Emission	
O	PCM linear	**E**	L2_256s	**D**	L2_384s	**E**	L2_256s	**F**	L2_256s-12
B	MP3_128js	**M**	L3_128js					**H**	L2_192js-12
C	MiniDisc	**P**	ADPCM_256s					**J**	L2_128js-12
W	WM_128s	**S**	AAC_128s						

4. Test sequences

Nine test sequences were initially chosen to represent a sufficiently critical range of broadcast material. Speech, music and single instruments are all covered in this selection. The sequences are all stereo, recorded at 48 kHz sampling frequency and range from 10 to 20 seconds duration. They are listed in Table 3.

[1] Except ADPCM_256s, which uses a 32 kHz sampling frequency

TABLE 3: Test items.

Name	Description	Origin
Accordion	Solo accordion music.	Swedish Radio
Castanets	Castanets.	EBU SQAM CD
Classic	Brass band music.	IRT
Dialog	German male and female conversation.	T-Systems
Harpsichord	Harpsichord playing an arpeggio.	EBU SQAM CD
Orchestra	Classical music.	IRT
Rea	Chris Rea.	Commercial CD
Vega	Suzanne Vega, "Tom's Diner" a cappella.	Commercial CD
Hockey	Commentary from ice hockey arena with crowd noise.	IRT

5. Overview of the test methodology

The test methodology consisted of three stages; the first two stages were intended to select the cascades used for the final, subjective, test stage.

5.1. Stage 1: Initial combination selection

The most common combinations were chosen from the 48 possibilities. It was also decided that with a Windows Media source the only likely contribution codec would be the Layer II 256 kbit/s. This reduced the number of combinations to 39.

5.2. Stage 2: Objective tests

The BBC and Radio France performed objective tests to obtain an initial evaluation of the quality of the different cascades and to reduce the combinations still further. The PEAQ objective test software was used on the 39 combinations to generate objective difference grade (ODG) scores for each of them.

It was decided that any combination scoring better than -1.0 was of sufficiently high quality and could be omitted from further testing. Of the 39 combinations tested, 3 items scored better than -1.0, leaving 36 combinations for subsequent testing.

5.3. Stage 3: Subjective tests

Subjective listening tests were performed using the MUSHRA methodology on the 36 combinations selected after the first two stages.

6. Objective tests (PEAQ)

Automated objective testing was a fast and efficient way of obtaining quality scores for all combinations of cascades using the nine chosen test sequences. The software used was the PEAQ algorithm, which generates ODGs ranging in value from -4 (very annoying) to 0 (indistinguishable from the original). From the ODGs generated it was possible to make a decision on the most critical combinations required for the subjective listening tests.

The scores agreed upon by the BBC and Radio France are shown in Table 4, where each test item (accordion (1) … hockey (9)) is scored for each cascade. The average for each cascade over all 9 items is shown, and this is how the table is ordered. The average score for each item over all cascades is shown at the bottom.

TABLE 4: PEAQ scores for cascades selected in first stage, for 9 items

Cascade	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Average
OPDEF	-0.594	-0.657	-0.468	-0.143	-1.411	-0.614	-0.506	-0.573	-0.178	-0.572
CPDEF	-0.920	-0.935	-0.728	-0.191	-2.053	-0.815	-0.672	-0.763	-0.334	-0.823
OEDEF	-0.869	-1.096	-0.679	-0.197	-1.783	-0.865	-0.795	-0.814	-0.403	-0.834
CEDEF	-1.192	-1.254	-0.889	-0.261	-2.338	-1.037	-0.927	-0.995	-0.521	-1.046
OPDEH	-1.296	-1.270	-1.146	-0.396	-2.587	-1.369	-1.121	-1.291	-0.569	-1.227
OEDEH	-1.567	-1.601	-1.324	-0.440	-2.779	-1.578	-1.349	-1.460	-0.784	-1.431
CPDEH	-1.640	-1.589	-1.350	-0.503	-2.920	-1.525	-1.267	-1.501	-0.769	-1.451
OSDEF	-1.887	-1.790	-1.434	-0.864	-2.645	-1.506	-1.644	-1.367	-1.155	-1.588
WEDEF	-1.393	-2.261	-1.583	-0.726	-2.435	-1.460	-1.679	-1.771	-1.258	-1.618
BPDEF	-1.670	-2.010	-1.587	-0.759	-2.696	-1.637	-1.762	-1.319	-1.231	-1.630
CEDEH	-1.862	-1.801	-1.909	-0.550	-3.032	-1.710	-1.473	-1.672	-0.910	-1.658
CSDEF	-1.744	-1.897	-1.567	-0.947	-2.908	-1.558	-1.720	-1.602	-1.283	-1.692
OMDEF	-1.611	-1.845	-1.621	-1.257	-2.977	-1.684	-1.784	-1.559	-1.241	-1.731
BEDEF	-1.840	-2.181	-1.648	-0.799	-2.868	-1.859	-1.913	-1.451	-1.314	-1.764
CMDEF	-1.888	-2.069	-1.768	-1.316	-3.189	-1.806	-1.949	-1.732	-1.401	-1.902
OSDEH	-2.097	-2.324	-2.137	-1.129	-3.189	-2.131	-2.177	-2.020	-1.548	-2.084
WEDEH	-2.018	-2.669	-2.090	-1.087	-3.117	-2.058	-2.315	-2.300	-1.639	-2.144
BPDEH	-2.243	-2.537	-2.115	-1.056	-3.259	-2.232	-2.267	-2.014	-1.612	-2.148
BSDEF	-2.175	-2.639	-2.011	-1.315	-3.179	-2.164	-2.471	-1.901	-1.733	-2.176
CSDEH	-2.296	-2.402	-2.104	-1.254	-3.325	-2.178	-2.355	-2.169	-1.678	-2.196
OMDEH	-2.193	-2.356	-2.158	-1.489	-3.367	-2.257	-2.259	-2.153	-1.669	-2.211
BEDEH	-2.365	-2.964	-2.151	-1.106	-3.304	-2.389	-2.417	-2.048	-1.668	-2.268
BMDEF	-2.330	-2.759	-2.147	-1.544	-3.348	-2.349	-2.562	-2.072	-1.851	-2.329
CMDEH	-2.387	-2.508	-2.259	-1.578	-3.458	-2.350	-2.382	-2.267	-1.817	-2.334
BSDEH	-2.616	-2.965	-2.448	-1.564	-3.468	-2.635	-2.767	-2.388	-2.019	-2.541
BMDEH	-2.717	-3.062	-2.527	-1.760	-3.558	-2.751	-2.982	-2.542	-2.110	-2.668
OPDEJ	-3.222	-3.127	-2.918	-2.091	-3.552	-3.271	-2.844	-2.948	-1.937	-2.879
OEDEJ	-3.261	-3.298	-2.947	-2.132	-3.569	-3.297	-2.757	-3.043	-2.034	-2.926
CPDEJ	-3.303	-3.228	-2.976	-2.149	-3.604	-3.311	-3.043	-3.016	-2.057	-2.965
CEDEJ	-3.320	-3.353	-2.979	-2.185	-3.633	-3.338	-2.813	-3.092	-2.098	-2.979
BPDEJ	-3.417	-3.532	-3.087	-2.453	-3.676	-3.457	-3.203	-3.169	-2.502	-3.166
BEDEJ	-3.433	-3.527	-3.089	-2.455	-3.685	-3.472	-3.370	-3.214	-2.476	-3.191
OSDEJ	-3.473	-3.469	-3.141	-2.452	-3.671	-3.505	-3.320	-3.259	-2.433	-3.192
OMDEJ	-3.427	-3.483	-3.149	-2.561	-3.704	-3.476	-3.243	-3.224	-2.485	-3.195
WEDEJ	-3.371	-3.655	-3.141	-2.504	-3.641	-3.440	-3.386	-3.294	-2.444	-3.208
CSDEJ	-3.428	-3.512	-3.133	-2.572	-3.693	-3.470	-3.348	-3.280	-2.460	-3.211
CMDEJ	-3.466	-3.516	-3.162	-2.632	-3.723	-3.491	-3.255	-3.265	-2.584	-3.233
BSDEJ	-3.508	-3.632	-3.180	-2.668	-3.722	-3.539	-3.349	-3.342	-2.685	-3.292
BMDEJ	-3.528	-3.663	-3.215	-2.780	-3.748	-3.575	-3.507	-3.338	-2.672	-3.336
Average	**-2.348**	**-2.524**	**-2.153**	**-1.432**	**-3.149**	**-2.337**	**-2.281**	**-2.185**	**-1.630**	**-2.227**

NOTE: The shaded cascades are those with an ODG better than -1.0.These were not subject to further testing. However, for item 5 (harpsichord), the scores for these three cascades were significantly worse than -1.0.

7. Cascades for Subjective Tests

Following the PEAQ tests the number of cascades to be subjectively tested was reduced to 36. See Table 5 below.

TABLE 5: The 36 cascades chosen for the EBU subjective tests.

CASCADE No.	CASCADE LABEL	INPUT FORMAT	CONTRIBUTION	STUDIO	DISTRIB	EMISSION
1	BEDEF	MP3_128js	L2_256s	L2_384s	L2_256s	L2_256s-12
2	BEDEH	MP3_128js	L2_256s	L2_384s	L2_256s	L2_192js-12
3	BEDEJ	MP3_128js	L2_256s	L2_384s	L2_256s	L2_128js-12
4	BMDEF	MP3_128js	L3_128js	L2_384s	L2_256s	L2_256s-12
5	BMDEH	MP3_128js	L3_128js	L2_384s	L2_256s	L2_192js-12
6	BMDEJ	MP3_128js	L3_128js	L2_384s	L2_256s	L2_128js-12
7	BPDEF	MP3_128js	ADPCM_256s	L2_384s	L2_256s	L2_256s-12
8	BPDEH	MP3_128js	ADPCM_256s	L2_384s	L2_256s	L2_192js-12
9	BPDEJ	MP3_128js	ADPCM_256s	L2_384s	L2_256s	L2_128js-12
10	BSDEF	MP3_128js	AAC_128s	L2_384s	L2_256s	L2_256s-12
11	BSDEH	MP3_128js	AAC_128s	L2_384s	L2_256s	L2_192js-12
12	BSDEJ	MP3_128js	AAC_128s	L2_384s	L2_256s	L2_128js-12
13	CEDEF	MiniDisc	L2_256s	L2_384s	L2_256s	L2_256s-12
14	CEDEH	MiniDisc	L2_256s	L2_384s	L2_256s	L2_192js-12
15	CEDEJ	MiniDisc	L2_256s	L2_384s	L2_256s	L2_128js-12
16	CMDEF	MiniDisc	L3_128js	L2_384s	L2_256s	L2_256s-12
17	CMDEH	MiniDisc	L3_128js	L2_384s	L2_256s	L2_192js-12
18	CMDEJ	MiniDisc	L3_128js	L2_384s	L2_256s	L2_128js-12
19	CPDEH	MiniDisc	ADPCM_256s	L2_384s	L2_256s	L2_192js-12
20	CPDEJ	MiniDisc	ADPCM_256s	L2_384s	L2_256s	L2_128js-12
21	CSDEF	MiniDisc	AAC_128s	L2_384s	L2_256s	L2_256s-12
22	CSDEH	MiniDisc	AAC_128s	L2_384s	L2_256s	L2_192js-12
23	CSDEJ	MiniDisc	AAC_128s	L2_384s	L2_256s	L2_128js-12
24	OEDEH	PCM linear	L2_256s	L2_384s	L2_256s	L2_192js-12
25	OEDEJ	PCM linear	L2_256s	L2_384s	L2_256s	L2_128js-12
26	OMDEF	PCM linear	L3_128js	L2_384s	L2_256s	L2_256s-12
27	OMDEH	PCM linear	L3_128js	L2_384s	L2_256s	L2_192js-12
28	OMDEJ	PCM linear	L3_128js	L2_384s	L2_256s	L2_128js-12
29	OPDEH	PCM linear	ADPCM_256s	L2_384s	L2_256s	L2_192js-12
30	OPDEJ	PCM linear	ADPCM_256s	L2_384s	L2_256s	L2_128js-12
31	OSDEF	PCM linear	AAC_128s	L2_384s	L2_256s	L2_256s-12
32	OSDEH	PCM linear	AAC_128s	L2_384s	L2_256s	L2_192js-12
33	OSDEJ	PCM linear	AAC_128s	L2_384s	L2_256s	L2_128js-12
34	WEDEF	WM_128s	L2_256s	L2_384s	L2_256s	L2_256s-12
35	WEDEH	WM_128s	L2_256s	L2_384s	L2_256s	L2_192js-12
36	WEDEJ	WM_128s	L2_256s	L2_384s	L2_256s	L2_128js-12

NOTE: The shaded cascades (BPDEJ, OSDEH and CEDEF) are those chosen as encompassing the entire quality dynamic range.

8. Sharing of Subjective Tests Workload

Subjective tests were carried out by the IRT, BBC R&D, the NRK and TVP

Table 6 shows how the 36 combinations were shared out amongst the laboratories. All laboratories tested three combinations (BPDEJ, OSDEH and CEDEF) to assess the correlation in scoring between the laboratories. NRK and TVP also had the OSDEH combination duplicated in their test ensemble due to the number of tests used.

TABLE 6: Allocation to the test laboratories of the cascades to be evaluated

1 IRT	2 BBC	3 TDA[1]	4 NRK	5 TVP
BPDEJ	BPDEJ	BPDEJ	BPDEJ	BPDEJ
OSDEH	OSDEH	OSDEH	OSDEH	OSDEH
CEDEF	CEDEF	CEDEF	CEDEF	CEDEF
BMDEJ	BSDEJ	CMDEJ	CSDEJ	WEDEJ
OMDEJ	OSDEJ	BEDEJ	CEDEJ	CPDEJ
OEDEJ	OPDEJ	BMDEH	BSDEH	CMDEH
BMDEF	BEDEH	OMDEH	CSDEH	BSDEF
BPDEH	WEDEH	CMDEF	BEDEF	OMDEF
CSDEF	CEDEH	BPDEF	WEDEF	OSDEF
CPDEH	OEDEH	OPDEH	OSDEH	OSDEH

The laboratories used the same subjective evaluation software, designed by Fraunhofer IDMT. The software was made available free of charge by the Fraunhofer Institute subject to signing a Non-Disclosure Agreement (NDA).

Each laboratory used at least 15 trained listeners – the number at each site is shown in the statistical analysis section. All laboratories used identical instructions (according to the MUSHRA specification given in ITU-R BS.1534).

9. Subjective Tests Methodology

The MUSHRA Methodology was used with the optional anchor of 10 kHz low-pass filtered.

9.1. Hidden reference and hidden anchors

A hidden reference and two hidden anchors, 3.5 kHz and 10 kHz low-pass filtered, were used. The higher anchor of 10 kHz was preferred to 7 kHz, as the subjective quality of the cascades was expected to be relatively high.

9.2. Preparation

Test items were prepared by the IRT with assistance from Radio France, and TVP. Commercially available codecs were used. For the studio codecs and the secondary distribution codecs hardware implementation was necessary. No frame alignment of samples was imposed.

9.3. Headphones

All tests were carried out using Stax SR-404 open-backed electrostatic headphones, with each laboratory using the same model. It was assumed the listening environment was quiet enough such that no ambient noise could interfere with the listening (e.g. ensuring noisy PCs were in a different room).

[1] Algerian EBU member TDA intended to perform listening tests also. Problems with Customs clearance meant that TDA had no access to necessary equipment and consequently the IRT and NRK shared this part of the work.

10. Statistical Analysis

Beate Klehs and Thomas Sporer of the Fraunhofer Institute (FhG) carried out the statistical analysis. Independent statistical analysis was performed for each laboratory. Based on the experience from the previous B/AIM tests, it was expected that the results from the different laboratories would be statistically coherent. Therefore the sharing of work could be done.

Thomas Sporer also performed the necessary analysis for the rejection of subjects (for each laboratory), as appropriate.

For the parametric statistics the mean was calculated with 95% confidence intervals.

10.1. Rejection of subjects

Some listening subjects who did not deliver consistent scores were rejected from the statistical analysis. The following criteria for rejection were used:

a) if a subject was consistently not able to discriminate between the (hidden) 10 kHz anchor and the (hidden) reference, he/she was rejected

b) if a subject consistently downgraded the reference by a significant amount (i.e. he/she was unable to detect it), he/she was rejected.

It was fortunate that few subjects were rejected, so that the number of "valid" subjects was large enough to perform a meaningful statistical analysis.

10.2. Number of subjects

The IRT used 18 subjects: 2 sets of results were removed and 16 were taken into account

The BBC used 21 subjects: 2 sets of results were removed and 19 were taken into account

The NRK used 18 subjects: 3 sets of results were removed and 15 were taken into account

The TVP used 15 subjects: 5 sets of results were removed and 10 were taken into account

11. Lessons Learned from the Test Process

Following these initial tests, the group felt it useful to collect experiences and draw some conclusions.

- Automated objective testing requires particular care to ensure that the test sequences are each properly synchronised with their respective reference signal. Sources of error included sample rate converters and free-running analogue to digital converters. The objective test software included some facilities to measure and track varying time-offsets, but this was found not to work reliably, particularly when the cascades caused significant impairment;
- A significant amount of time must be spent on training in order to allow subjects to become familiar with test sequences, the artefacts and the user interface;
- A very low background noise level in the room is essential even if headphones are being used since these do not necessarily block out sound;
- The subjects should listen through the whole duration of all test items and not make their assessment only of the beginning. The design of the subjective test software can influence this behaviour;
- Subjects above the age of 50 may be less able to distinguish the reference and the 10 kHz anchor due to a decrease in sensitivity to high frequencies. In general somewhat younger subjects with good auditory capabilities should be used;

- The sessions must be short, with adequate breaks between sessions, so that the subjects do not lose their concentration and fail to notice artefacts they would normally identify or mis-manipulate the assessment sliders. The design of the user interface plays an increasingly important role as the tests become more arduous. Much can be done to reduce errors made by subjects if clear on-screen indications are given;

The analysis of the results shows that the subjects may fall into one of the following categories:

- "Experienced" subject: able to discriminate different qualities; uses the whole scale, according to the quality detected;
- "Overly critical" subject: scores more often lower scores; may be too critical or too dismissive. However, it is not possible to dismiss a person's honestly held opinion in a subjective test;
- "Prudent" subject: scores mostly in the middle of the scale, avoiding extreme values (0 and 100). With training (and experience) this listener can become "experienced";
- "Unreliable" subject: is unable to hear artefacts; gives uncritically high scores; sometimes gives low scores to unimpaired items.

12. Evaluation Results

Table 7 lists all 36 cascades used in the subjective tests, and gives the average scores over all sites and test items. The 95% confidence intervals are also listed.

TABLE 7: Final results of subjective evaluations of 36 cascades

STUDIO: L2_384s

DISTRIBUTION: L2_256s

No.	CODE	INPUT FORMAT	CONTRIBUTION	EMISSION	SUBJECTIVE QUALITY	
					Average[1]	Conf. Interval
1	BEDEF	MP3_128js	L2_256s	L2_256s-12	69.51	3.92
2	BEDEH	MP3_128js	L2_256s	L2_192js-12	67.38	3.76
3	BEDEJ	MP3_128js	L2_256s	L2_128js-12	48.61	4.75
4	BMDEF	MP3_128js	L3_128js	L2_256s-12	63.38	4.44
5	BMDEH	MP3_128js	L3_128js	L2_192js-12	62.02	4.60
6	BMDEJ	MP3_128js	L3_128js	L2_128js-12	38.83	4.47
7	BPDEF	MP3_128js	ADPCM_256s	L2_256s-12	76.51	3.88
8	BPDEH	MP3_128js	ADPCM_256s	L2_192js-12	63.90	4.37
9	BPDEJ	MP3_128js	ADPCM_256s	L2_128js-12	49.36	2.02
10	BSDEF	MP3_128js	AAC_128s	L2_256s-12	76.83	3.43
11	BSDEH	MP3_128js	AAC_128s	L2_192js-12	61.52	4.31
12	BSDEJ	MP3_128js	AAC_128s	L2_128js-12	46.103	4.34
13	CEDEF	MiniDisc	L2_256s	L2_256s-12	82.20	1.43
14	CEDEH	MiniDisc	L2_256s	L2_192js-12	77.09	3.30
15	CEDEJ	MiniDisc	L2_256s	L2_128js-12	50.56	4.14
16	CMDEF	MiniDisc	L3_128js	L2_256s-12	74.48	4.39
17	CMDEH	MiniDisc	L3_128js	L2_192js-12	74.21	3.85
18	CMDEJ	MiniDisc	L3_128js	L2_128js-12	50.02	4.72
19	CPDEH	MiniDisc	ADPCM_256s	L2_192js-12	72.25	4.03
20	CPDEJ	MiniDisc	ADPCM_256s	L2_128js-12	66.23	4.29

[1] Across all 9 items and, for common cascades, across 3 laboratories

No.	CODE	INPUT FORMAT	CONTRIBUTION	EMISSION	SUBJECTIVE QUALITY	
					Average[1]	Conf. Interval
21	CSDEF	MiniDisc	AAC_128s	L2_256s-12	72.30	3.79
22	CSDEH	MiniDisc	AAC_128s	L2_192js-12	67.80	4.11
23	CSDEJ	MiniDisc	AAC_128s	L2_128js-12	50.10	4.58
24	OEDEH	PCM linear	L2_256s	L2_192js-12	79.07	3.27
25	OEDEJ	PCM linear	L2_256s	L2_128js-12	46.33	4.59
26	OMDEF	PCM linear	L3_128js	L2_256s-12	80.81	3.05
27	OMDEH	PCM linear	L3_128js	L2_192js-12	70.77	4.47
28	OMDEJ	PCM linear	L3_128js	L2_128js-12	43.33	4.62
29	OPDEH	PCM linear	ADPCM_256s	L2_192js-12	78.42	4.14
30	OPDEJ	PCM linear	ADPCM_256s	L2_128js-12	56.58	4.32
31	OSDEF	PCM linear	AAC_128s	L2_256s-12	85.65	2.46
32	OSDEH	PCM linear	AAC_128s	L2_192js-12	71.18	1.62
33	OSDEJ	PCM linear	AAC_128s	L2_128js-12	49.12	4.24
34	WEDEF	WM_128s	L2_256s	L2_256s-12	80.17	3.16
35	WEDEH	WM_128s	L2_256s	L2_192js-12	73.67	3.39
36	WEDEJ	WM_128s	L2_256s	L2_128js-12	64	4.32

From this table it can be seen that there is a wide range of quality. Some obvious advice can be derived:

Cascades to be avoided (mean opinion score less than 60) -

BMDEJ, BSDEJ, CMDEJ, CSDEJ, OMDEJ, OSDEJ, BEDEJ, BPDEJ, CEDEJ, OEDEJ, OPDEJ.

All these cascades have 128 kbit/s Layer II as the emission codec, which implies that a bitrate of 128 kbit/s is too low. However, this choice is often beyond the influence of the production engineer.

Cascades that cause the least degradation (mean opinion score more than 80) -

CEDEF, OMDEF, OSDEF, WEDEF, together with OPDEF, CPDEF, and OEDEF based on their objective test result.

All these cascades have 256 kbit/s Layer II as the emission codec, but it is not always possible to use such a high bit-rate if the broadcaster has chosen to maximise the number of services carried on a multiplex.

Four of these cascades had linear PCM (O) as the input ‘codec’, which does demonstrate the importance of keeping audio signals “uncoded” and reducing the occurrence of cascading as much as possible.

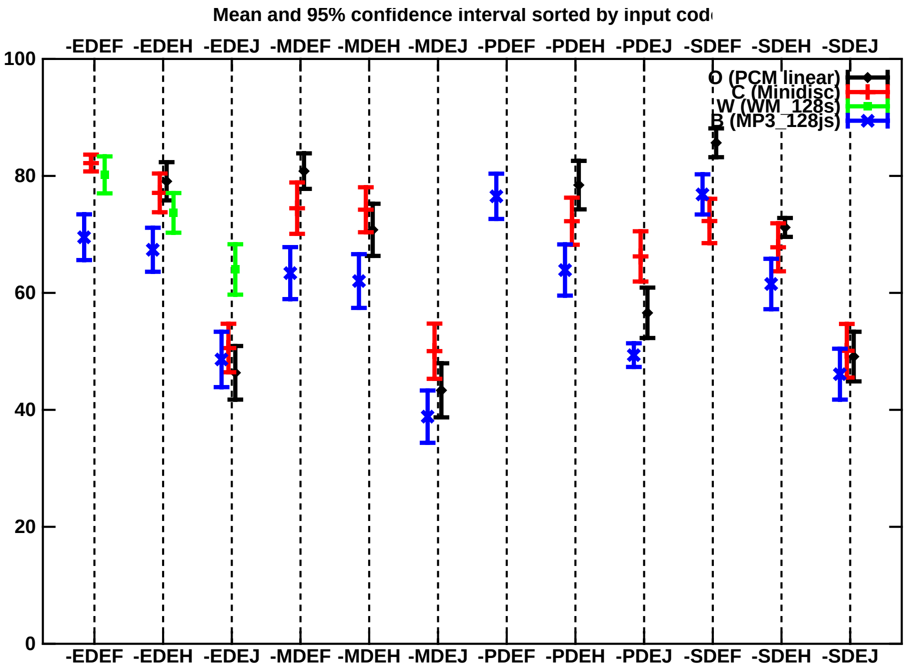
None of the cascades performed sufficiently well to recommend them without some reservation. The objective tests revealed that for item 5 (harpsichord) the best score achieved was -1.411, with the majority of the cascades scoring worse than -2.0. Therefore for some material even the best cascades can produce noticeably degraded audio.

12.1. Cascade performance as a function of input, contribution, and emission codecs

To assess the performance of each codec in the chain with respect to the others in the various cascades, graphs, with each of the 11 codecs kept constant, have been produced. There are three graphs, showing performance according to input codec, contribution codec and emission codec. The results shown are the average across all test items.

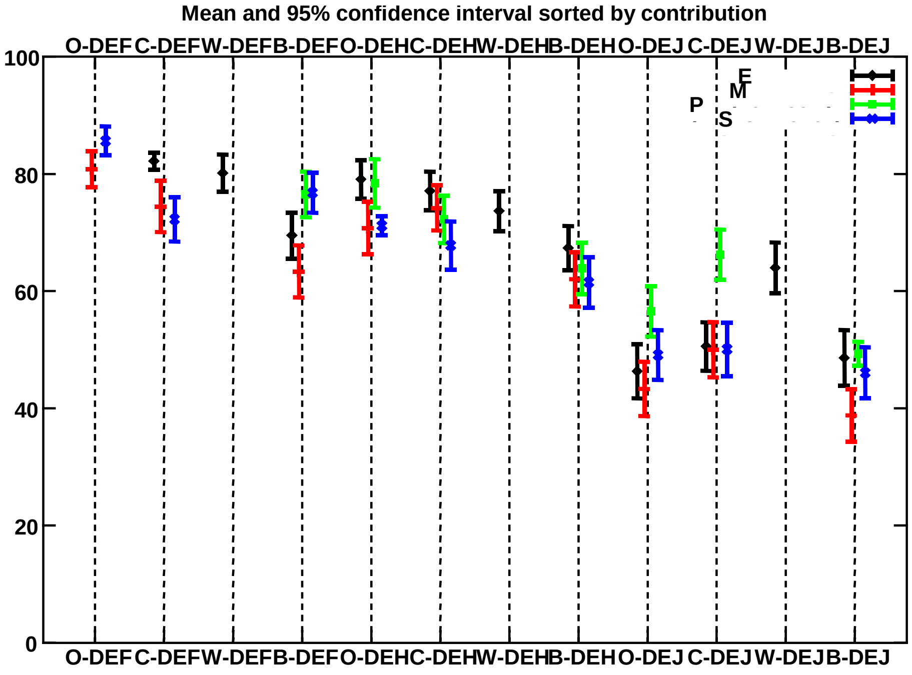
12.1.1. Results grouped according to input codec

The four input codecs, PCM linear (O), MP3 at 128 kbit/s joint-stereo (B), Minidisc (C) and Windows Media 128 kbit/s (W), are shown in the graph below. Windows media has only three scores, as it was only tested with one contribution codec.



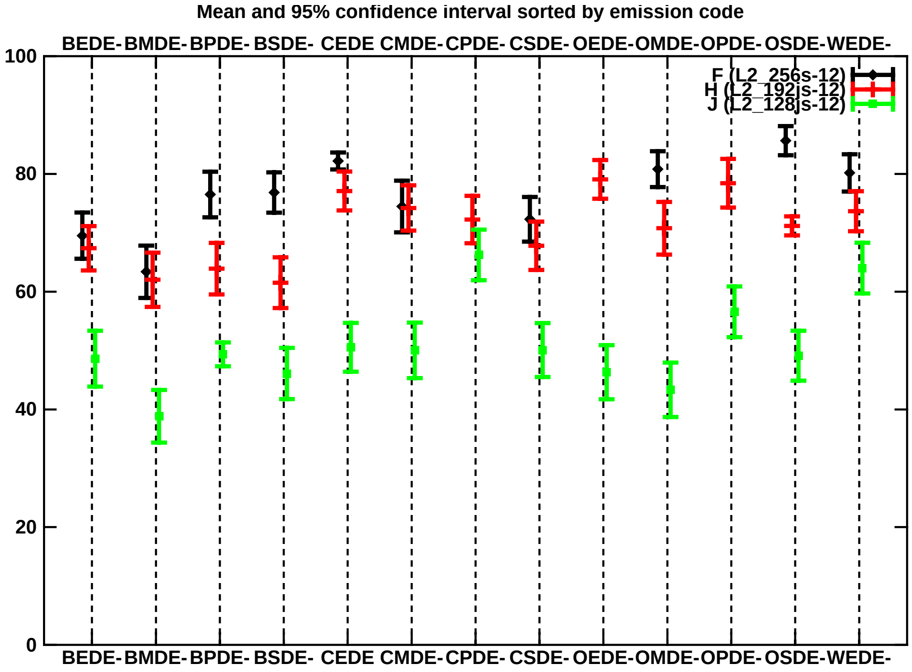
12.1.2. Results grouped according to contribution codec

The four contribution codecs, Layer II at 256 kbit/s stereo (E), Layer II at 128 kbit/s joint-stereo (M), ADPCM at 256 kbit/s stereo (P) and AAC at 128 kbit/s stereo (S) are shown in the graph below.



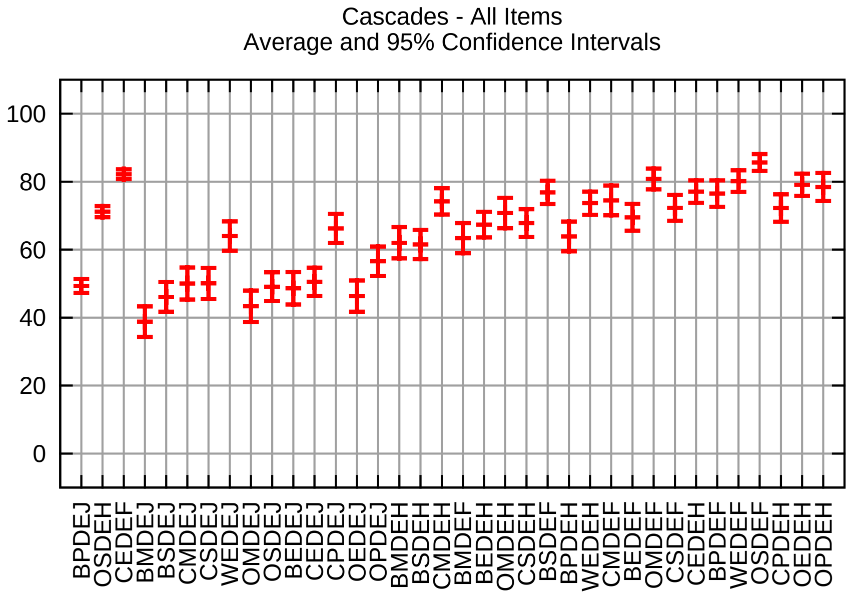
12.1.3. Results grouped according to emission codec

The three emission codecs, Layer II at 256 kbit/s stereo (F), Layer II at 192 kbit/s joint-stereo (H) and Layer II at 128 kbit/s joint-stereo (J) are shown in the graph below.



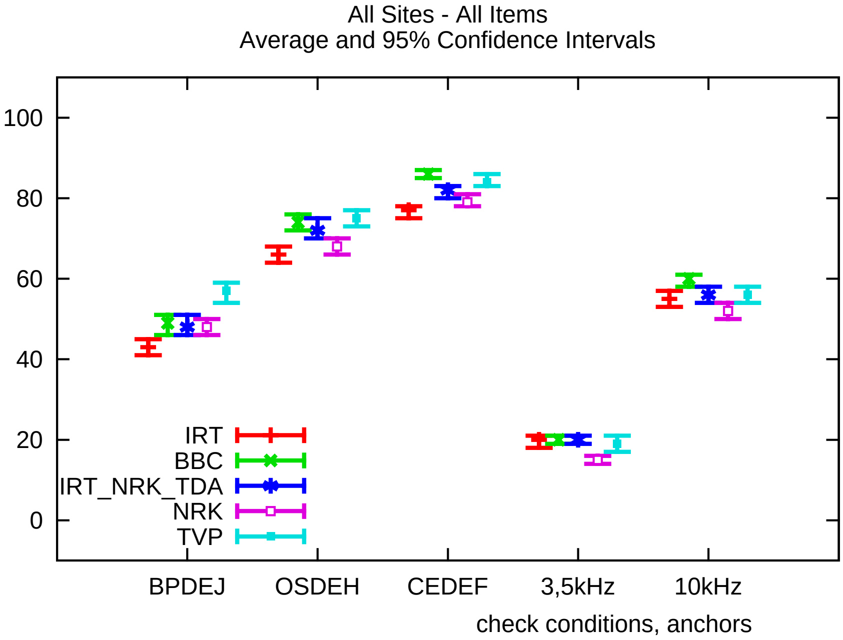
12.2. Overall graph of 36 cascades

All 36 cascades plotted in one graph, both the average and 95% confidence intervals over all items are shown below. The first three cascades are those that were tested by all the labs. The remaining cascades are ordered by their objective (PEAQ) test scores. The upward trend from left to right implies a good correlation between objective and subjective scores.



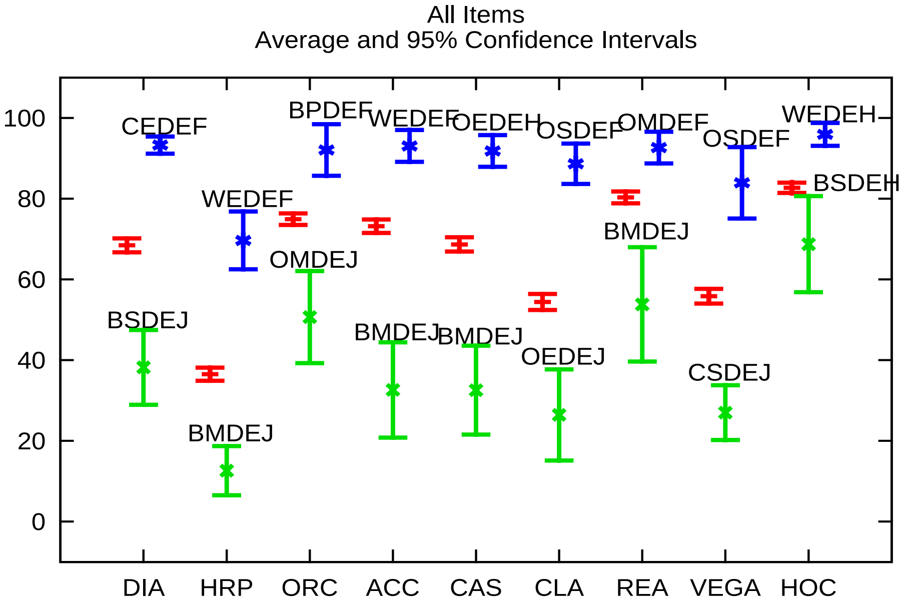
12.3. Comparing laboratories

Three cascades (BPDEJ, OSDEH and CEDEF) were tested by all the labs, which allowed a comparison between the scoring of the labs. The hope was that all the labs' results for these cascades would be very similar. The graph shows that on average the IRT scores are the lowest and that TVP's are the highest. These differences are likely to affect the scores for all the other cascades that were tested by one lab.



12.4. Comparing the performance of each test item over all cascades

The performance of the 9 test items was expected to vary greatly, for example the harpsichord is known to be particularly sensitive to audio coding. The average, minimum and maximum cascade scores for each item are shown in this graph.

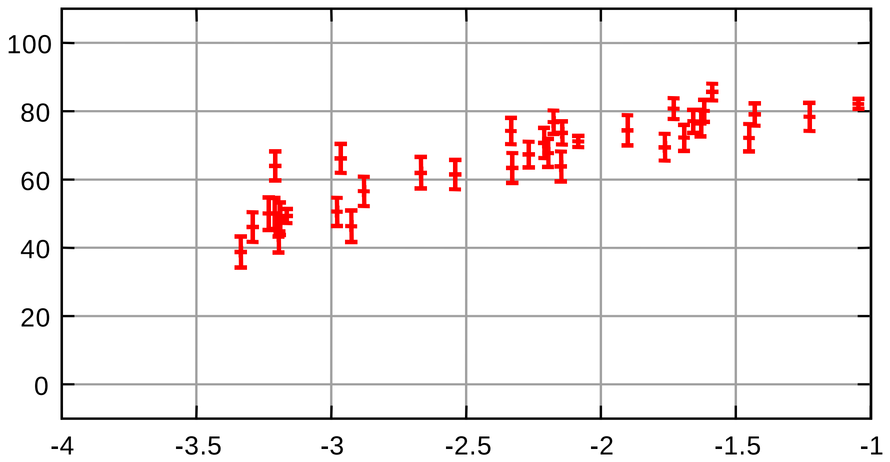


As expected the harpsichord (HRP) has the lowest average, minimum and maximum scores. The ice hockey (HOC) has the highest minimum and averages, which means that it was least sensitive to coding, and thus the least useful item for critical listening. The item with the greatest difference between minimum and maximum is the clarinet (CLA), so it is very well suited to critical listening.

12.5. Correlation of objective and subjective scores.

Objective measurements using the PEAQ software were taken as an initial guide to the quality of the cascades. The graph shown below compares these scores with those of the subjective test, to observe the amount of correlation. It must be noted that PEAQ gave impairment scores ranging from 0 to -4, whereas the subjective scoring with MUSHRA gave quality scores ranging from 100 to 0. While we can expect the two to be closely related, they are not, strictly speaking, measurements of the same thing on differing scales.

SDG - ODG -
Average and 95% Confidence



The graph shows that an ODG of -3 gives an SDG score of between 40 and 60. This corresponds to an objective impairment of “very annoying” and a subjective quality of “fair”. In a production chain any audio that is considered as “very annoying”, and therefore unlikely to be considered to be good enough for use, cannot seriously be classified as “fair”.

13. Conclusions

A thorough, extensive and time-consuming investigation has been conducted into cascaded audio coding. A model of a broadcast chain consisting of 5 cascaded codecs was assumed. From the thousands of possible combinations of codecs, a subset of the more likely ones was tested for audio performance using objective and subjective methods.

PEAQ objective testing was successfully employed to reduce the number of combinations that needed to be subjectively tested. The subjective testing was performed using the MUSHRA test method, with the subset of codec combinations being divided amongst a small number of test laboratories. Some codec cascades were tested by all sites for comparison purposes.

The results clearly show that the cumulative effect of cascaded audio coding can be highly detrimental to audio quality, even when each stage in the chain accounts for only a small reduction in quality.

The comparison of objective and subjective results showed a good correlation between scores. Caution should be exercised here because the scales and descriptive terms associated with the two test methods used are quite different.

The objective and subjective test results were both analysed to try to identify codec performance that was significantly better or significantly worse than expected. It was found that none of the combinations showed any unusual behaviour. This should simplify the selection process for users of low bit rate coding - it implies that choosing the best codecs will give better results. If the best possible quality is required, then coding must be avoided completely.

13.1. Caveats

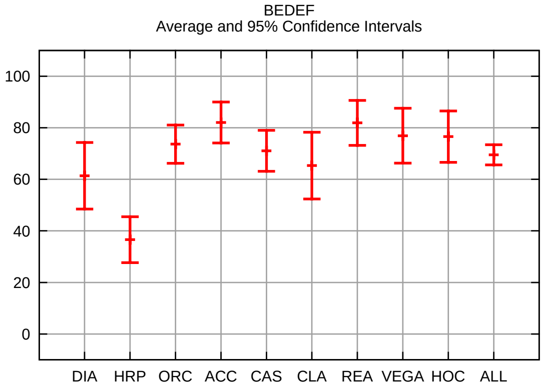
Dividing the subjective tests over several test sites revealed some rather troublesome side effects with the results.

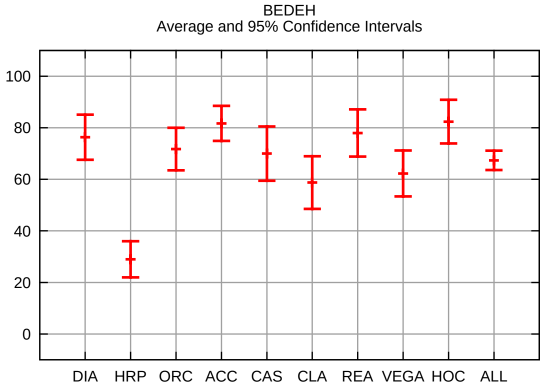
Not all the sites performed similarly enough to make the scores comparable. The three common cascades spread across all sites produced differing averages, and this pattern was reflected in the other cascade scores.

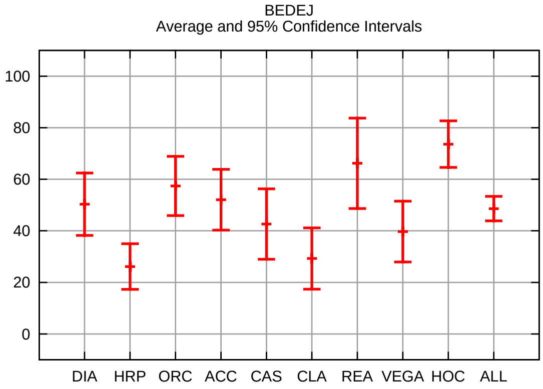
The issue was made more confusing by the performance of the two anchors being well matched between sites. This reveals the problem with using anchors that sound so different from the tested audio. Clearly some listeners do not mind coding artefacts as much as heavily band limited audio.

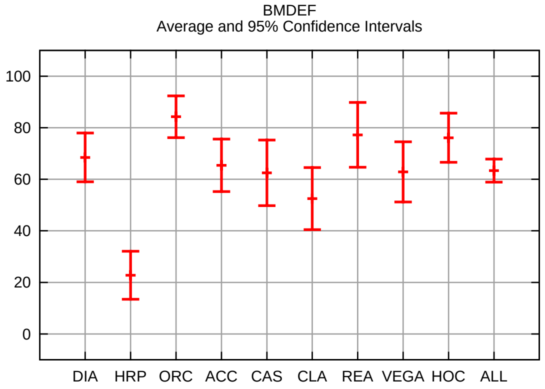
14. APPENDIX: Graphical Results

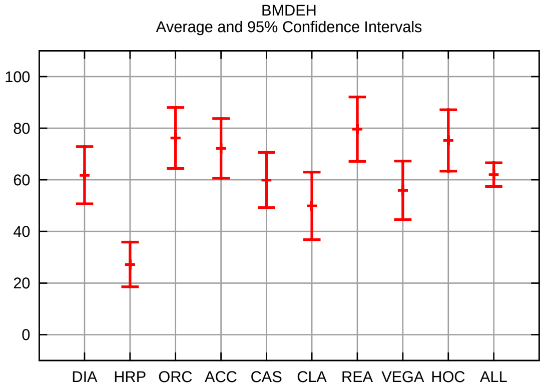
The following graphs show the scores for each cascade comparing each test item.

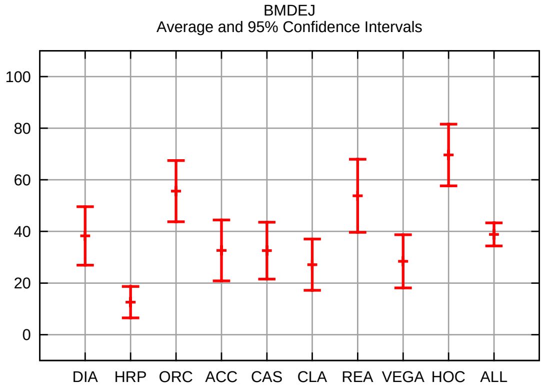




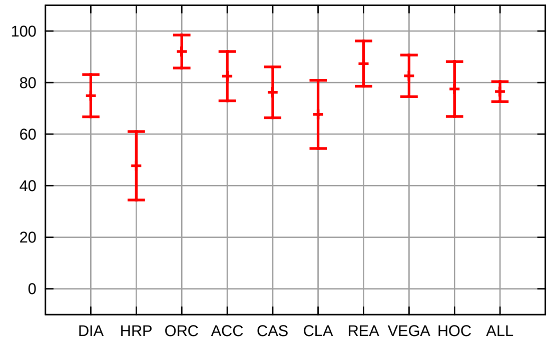




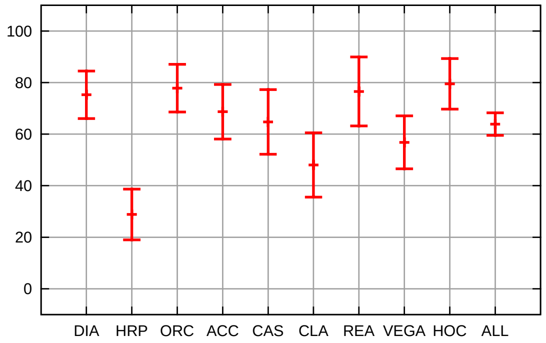




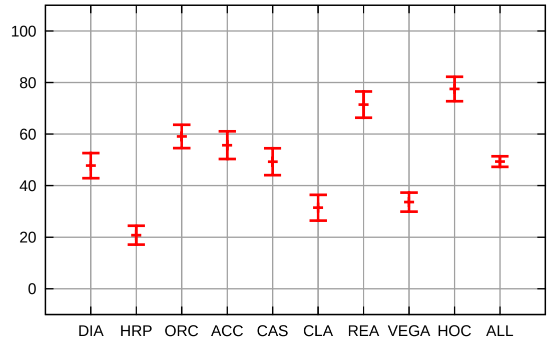
BPDEF
Average and 95% Confidence Intervals



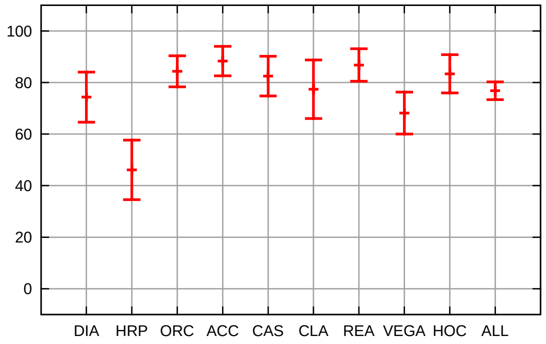
BPDEH
Average and 95% Confidence Intervals



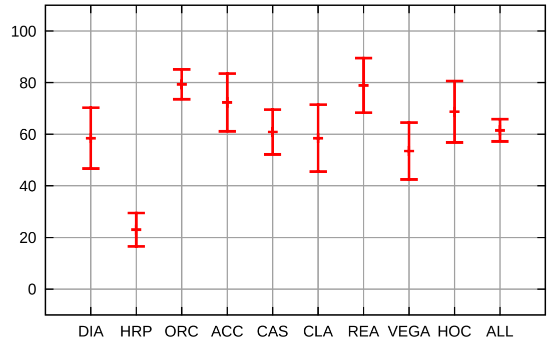
BPDEJ
Average and 95% Confidence Intervals



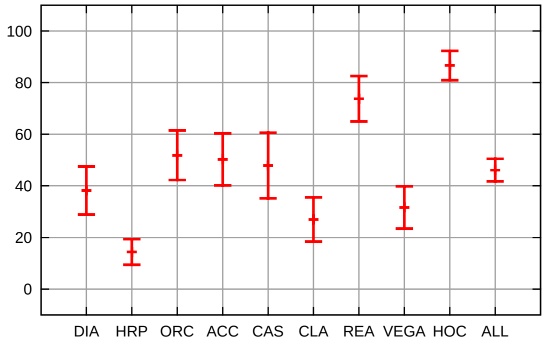
BSDEF
Average and 95% Confidence Intervals



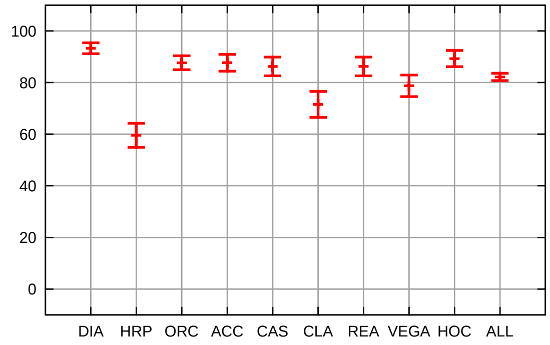
BSDEH
Average and 95% Confidence Intervals



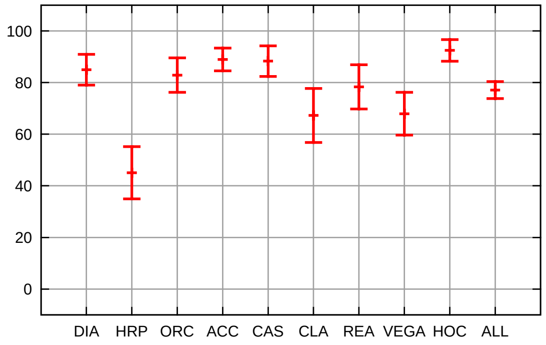
BSDEJ
Average and 95% Confidence Intervals



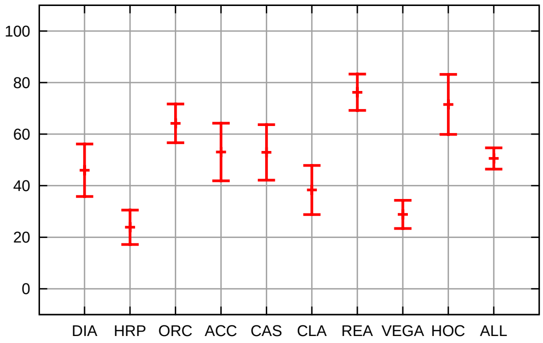
CEDEF
Average and 95% Confidence Intervals



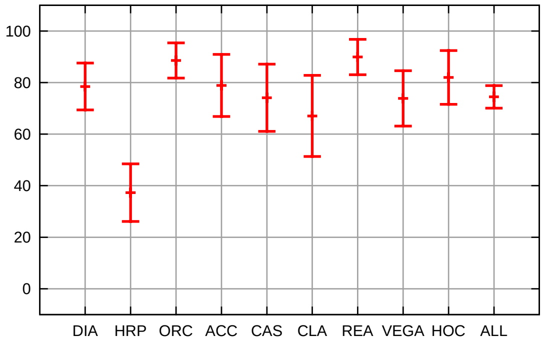
CEDEH
Average and 95% Confidence Intervals



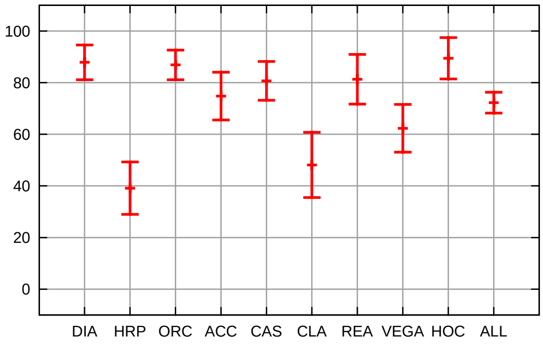
CEDEJ
Average and 95% Confidence Intervals



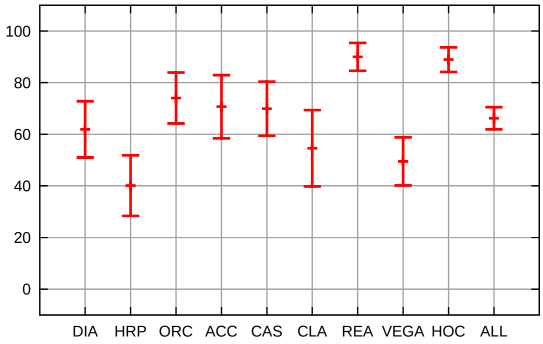
CMDEF
Average and 95% Confidence Intervals



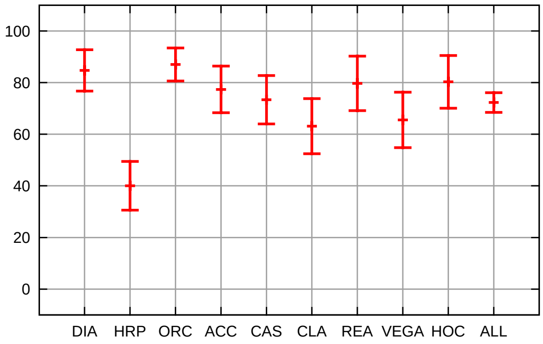
CPDEH
Average and 95% Confidence Intervals



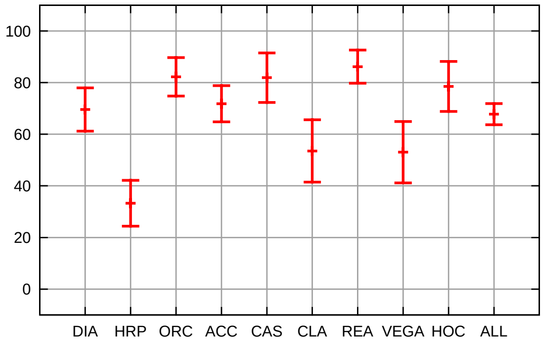
CPDEJ
Average and 95% Confidence Intervals



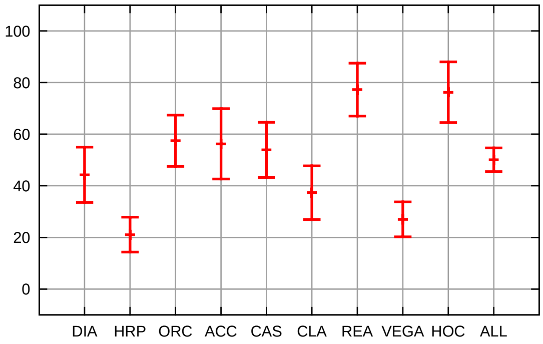
CSDEF
Average and 95% Confidence Intervals



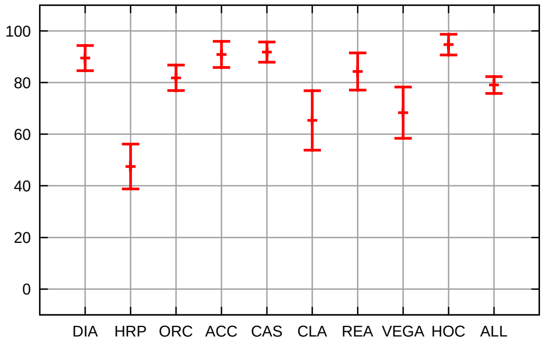
CSDEH
Average and 95% Confidence Intervals



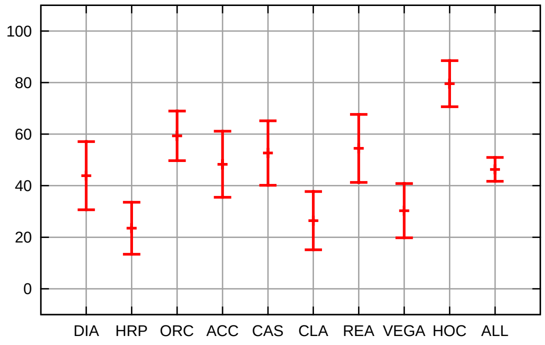
CSDEJ
Average and 95% Confidence Intervals



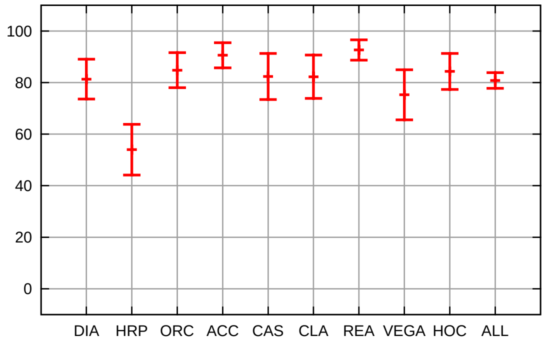
OEDEH
Average and 95% Confidence Intervals



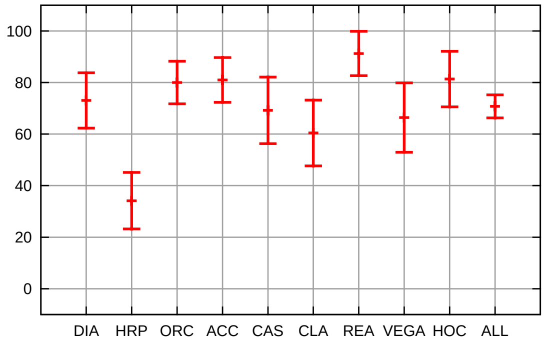
OEDEJ
Average and 95% Confidence Intervals



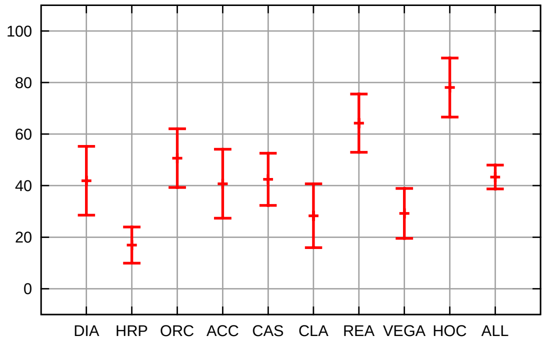
OMDEF
Average and 95% Confidence Intervals



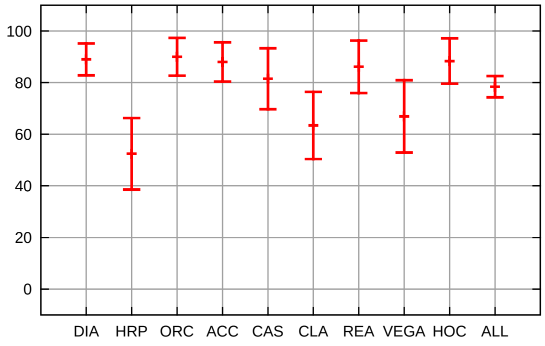
OMDEH
Average and 95% Confidence Intervals



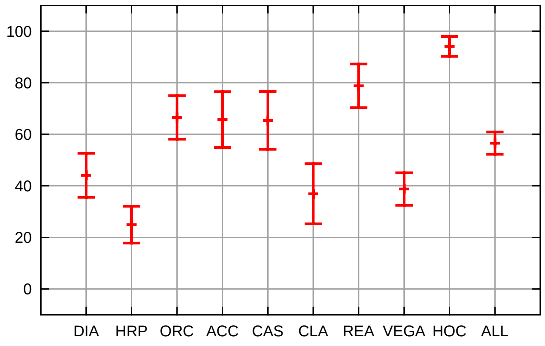
OMDEJ
Average and 95% Confidence Intervals



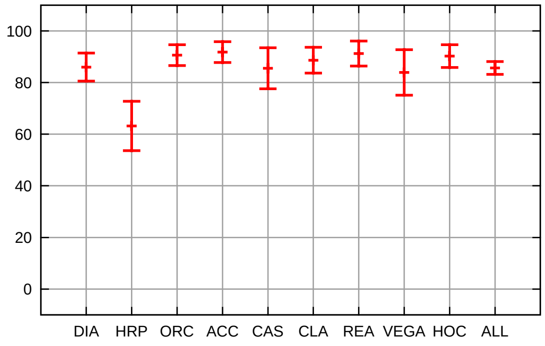
OPDEH
Average and 95% Confidence Intervals



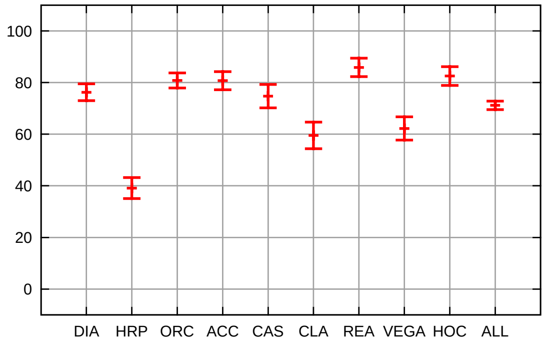
OPDEJ
Average and 95% Confidence Intervals



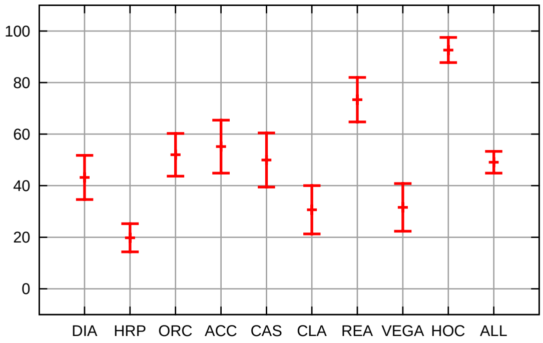
OSDEF
Average and 95% Confidence Intervals



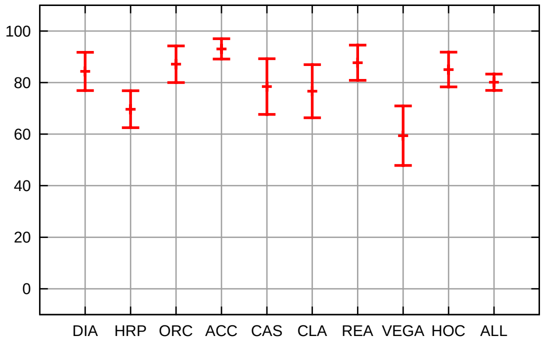
OSDEH
Average and 95% Confidence Intervals



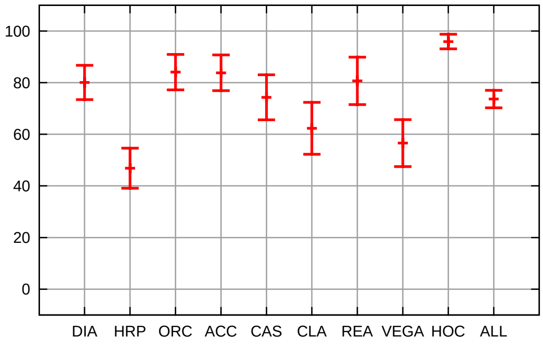
OSDEJ
Average and 95% Confidence Intervals



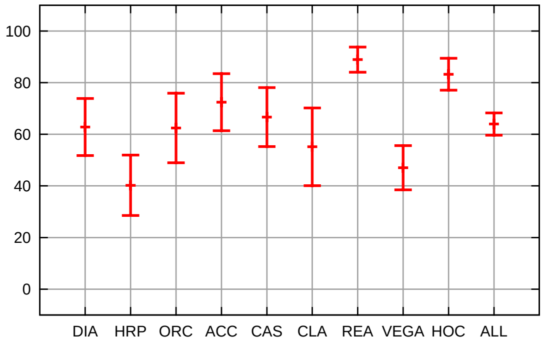
WEDEF
Average and 95% Confidence Intervals



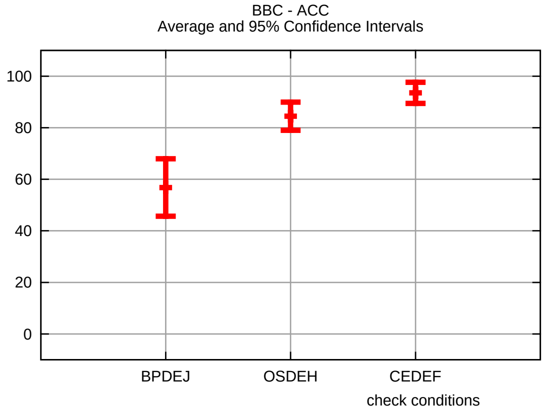
WEDEH
Average and 95% Confidence Intervals

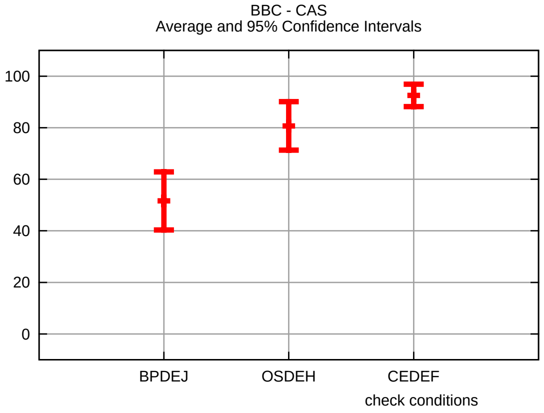


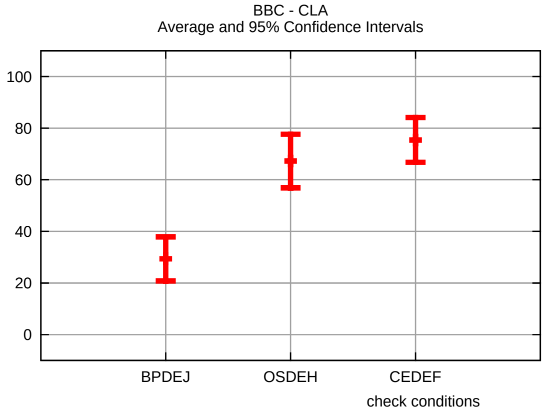
WEDEJ
Average and 95% Confidence Intervals

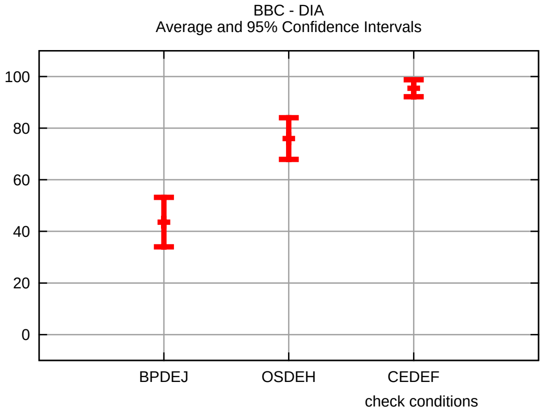


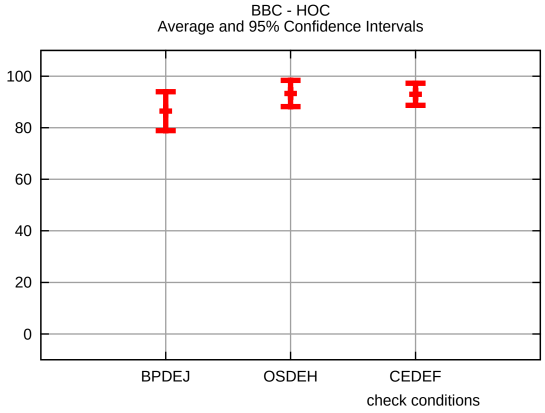
The following graphs show how each test site performed with the three common cascades for each test item.

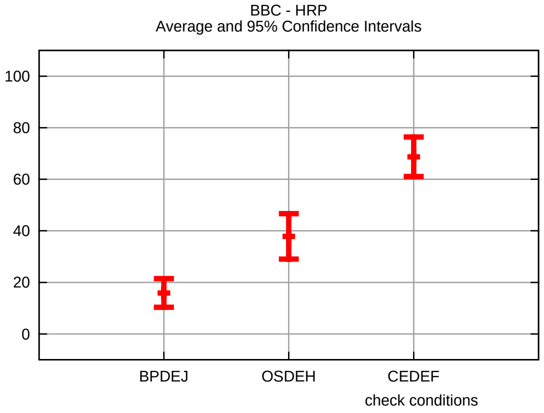


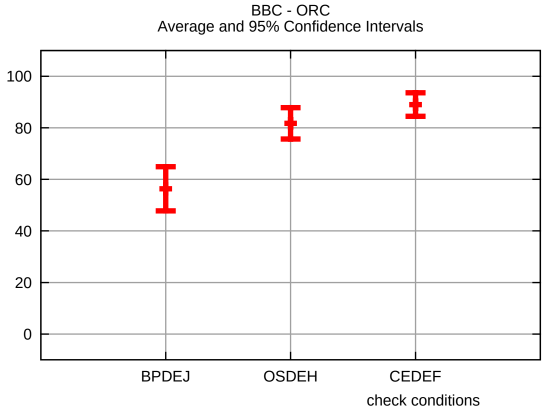


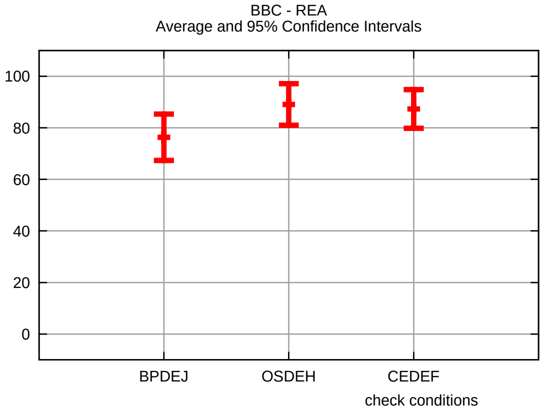


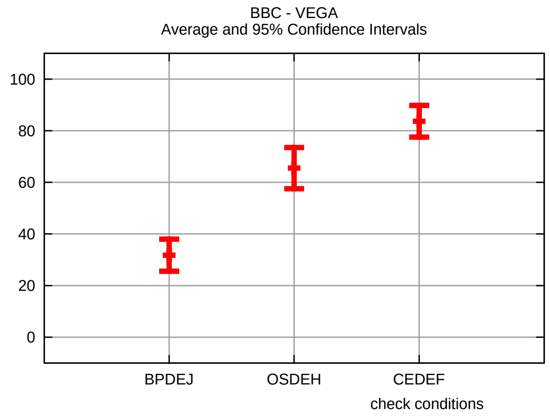


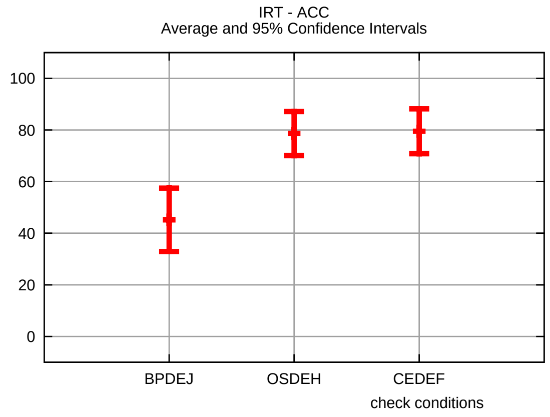


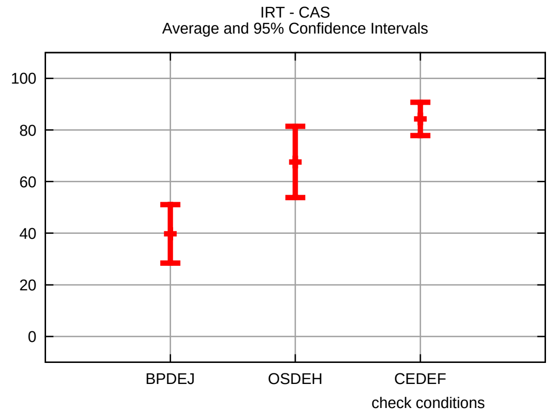


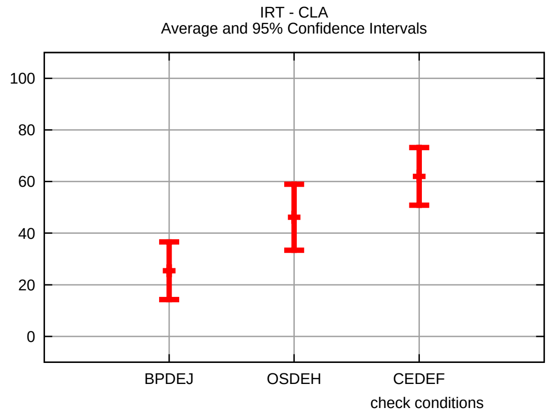


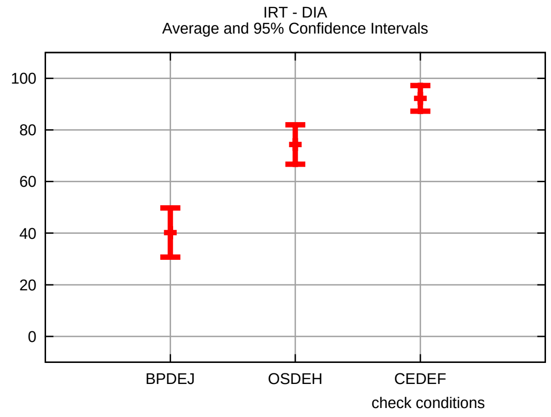


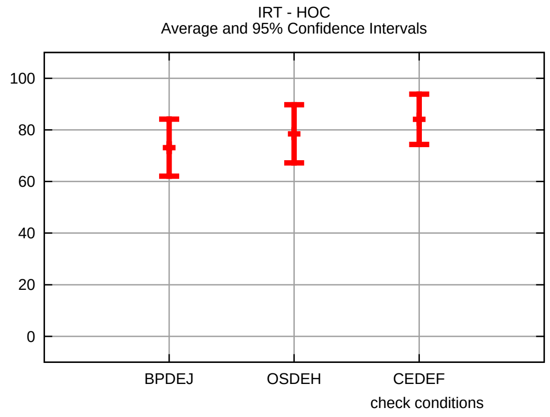

BBC - VEGA
Average and 95% Confidence Intervals
100
80
60
40
20
0
BPDEJ
OSDEH
CEDEF
check conditions

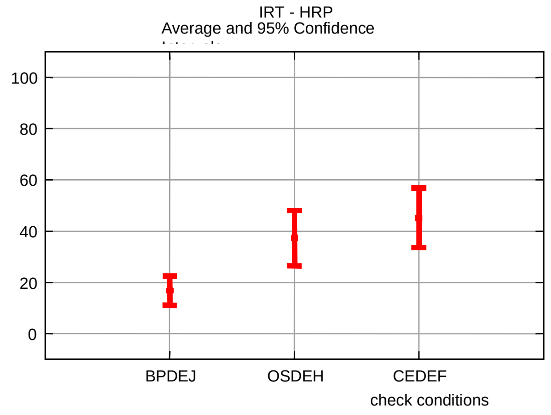

IRT - ACC
Average and 95% Confidence Intervals
100
80
60
40
20
0
BPDEJ
OSDEH
CEDEF
check conditions

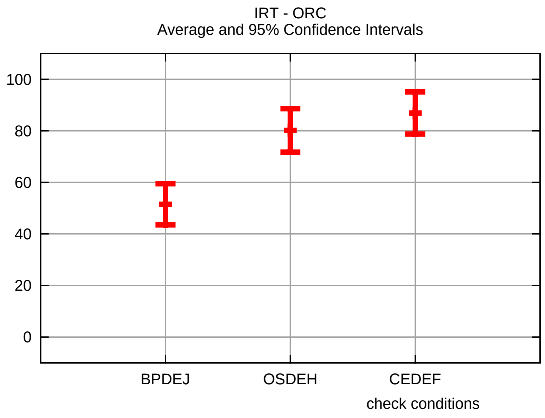

IRT - CAS
Average and 95% Confidence Intervals
100
80
60
40
20
0
BPDEJ
OSDEH
CEDEF
check conditions

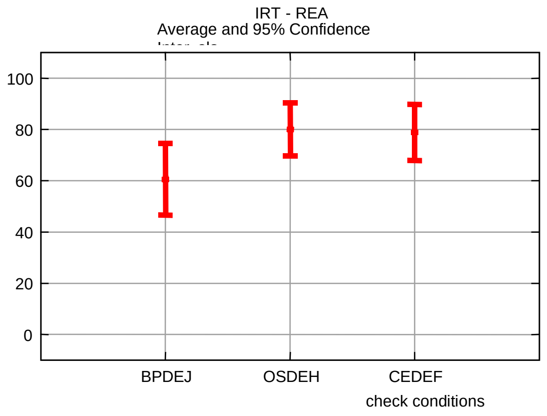

IRT - CLA
Average and 95% Confidence Intervals
100
80
60
40
20
0
BPDEJ
OSDEH
CEDEF
check conditions

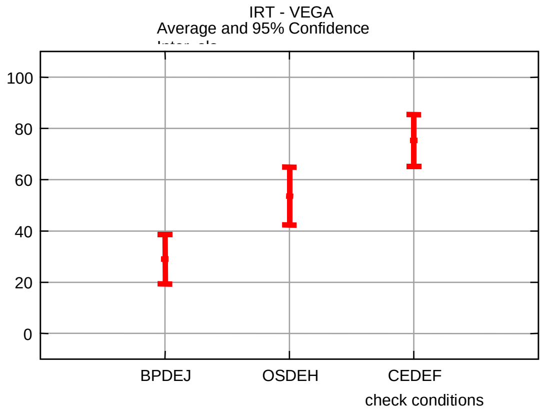

IRT - DIA
Average and 95% Confidence Intervals
100
80
60
40
20
0
BPDEJ
OSDEH
CEDEF
check conditions

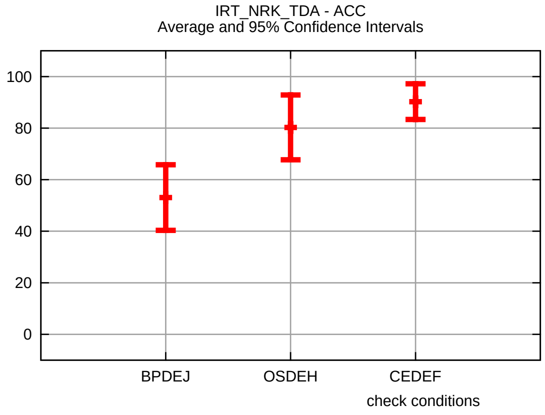

IRT - HOC
Average and 95% Confidence Intervals
100
80
60
40
20
0
BPDEJ
OSDEH
CEDEF
check conditions

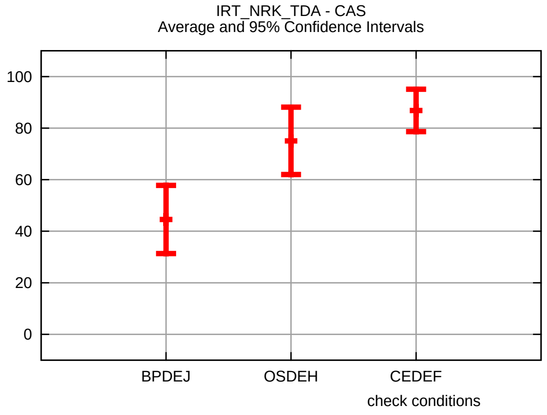

IRT - HRP
Average and 95% Confidence
100
80
60
40
20
0
BPDEJ
OSDEH
CEDEF
check conditions

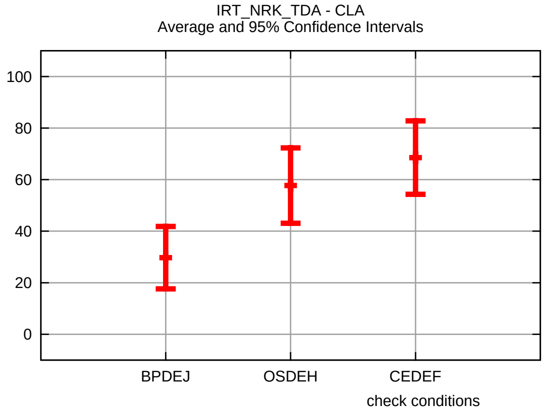

IRT - ORC
Average and 95% Confidence Intervals
100
80
60
40
20
0
BPDEJ
OSDEH
CEDEF
check conditions

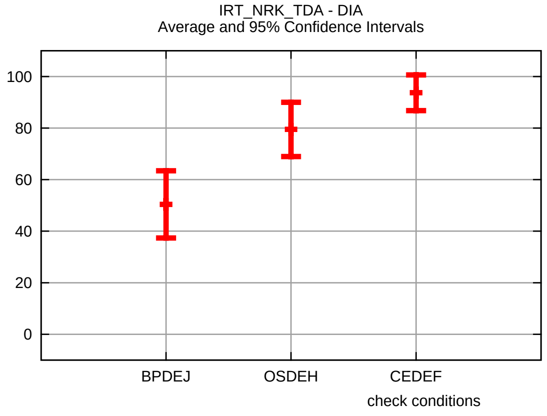

IRT - REA
Average and 95% Confidence
100
80
60
40
20
0
BPDEJ
OSDEH
CEDEF
check conditions


IRT - VEGA
Average and 95% Confidence
100
80
60
40
20
0
BPDEJ
OSDEH
CEDEF
check conditions

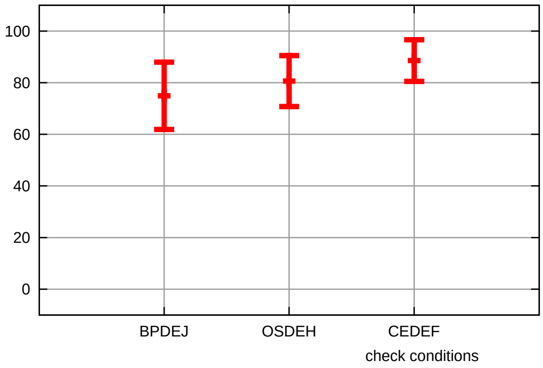

IRT_NRK_TDA - ACC
Average and 95% Confidence Intervals
100
80
60
40
20
0
BPDEJ
OSDEH
CEDEF
check conditions


IRT_NRK_TDA - CAS
Average and 95% Confidence Intervals
100
80
60
40
20
0
BPDEJ
OSDEH
CEDEF
check conditions

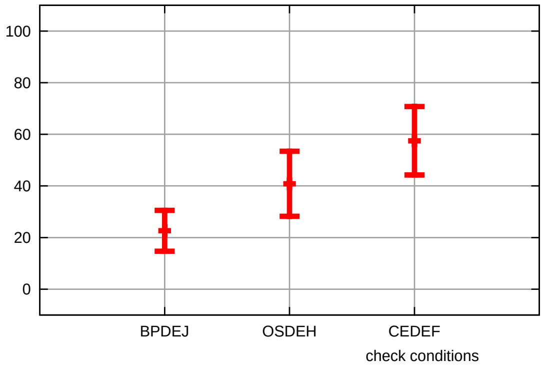

IRT_NRK_TDA - CLA
Average and 95% Confidence Intervals
100
80
60
40
20
0
BPDEJ
OSDEH
CEDEF
check conditions


IRT_NRK_TDA - DIA
Average and 95% Confidence Intervals
100
80
60
40
20
0
BPDEJ
OSDEH
CEDEF
check conditions

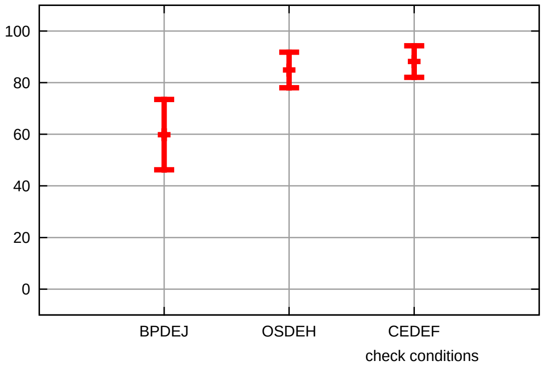
IRT_NRK_TDA - HOC
Average and 95% Confidence Intervals



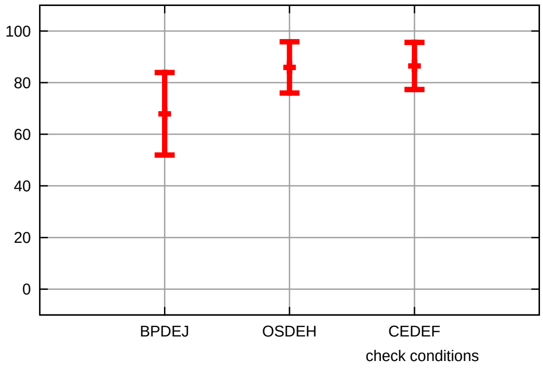
IRT_NRK_TDA - HRP
Average and 95% Confidence Intervals



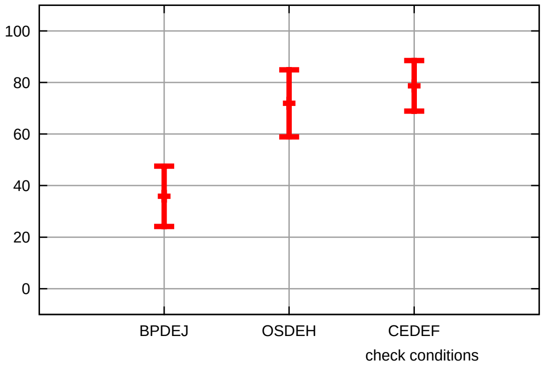
IRT_NRK_TDA - ORC
Average and 95% Confidence Intervals



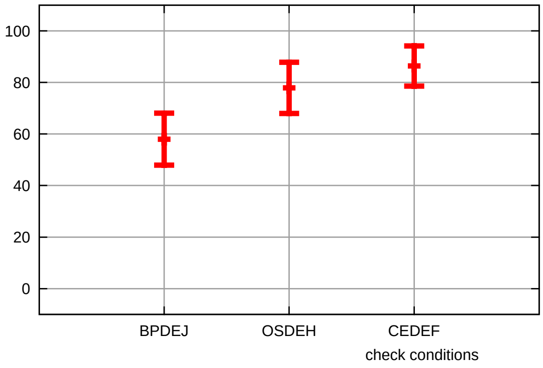
IRT_NRK_TDA - REA
Average and 95% Confidence Intervals



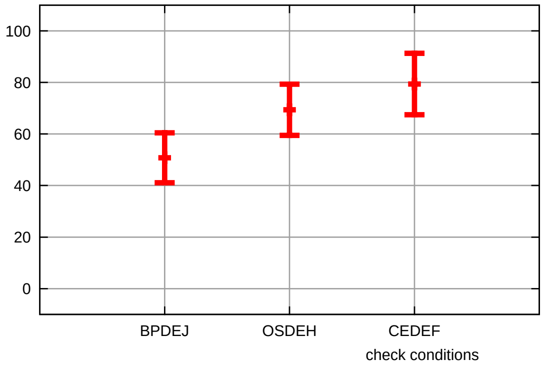
IRT_NRK_TDA - VEGA
Average and 95% Confidence Intervals



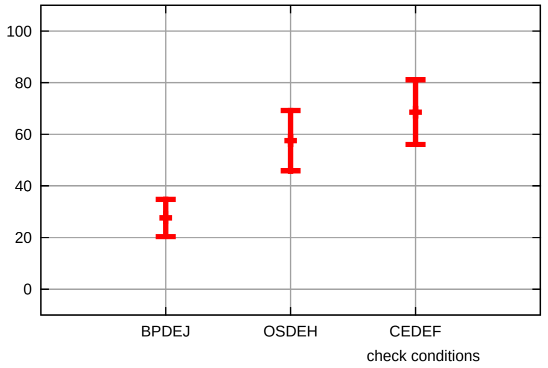
NRK - ACC
Average and 95% Confidence Intervals



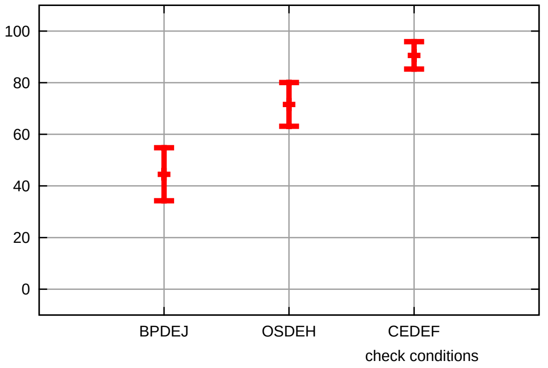
NRK - CAS
Average and 95% Confidence Intervals



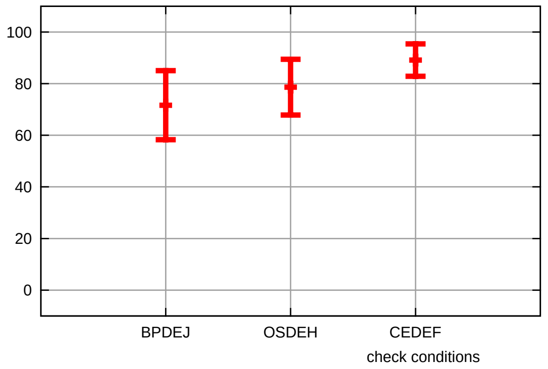
NRK - CLA
Average and 95% Confidence Intervals



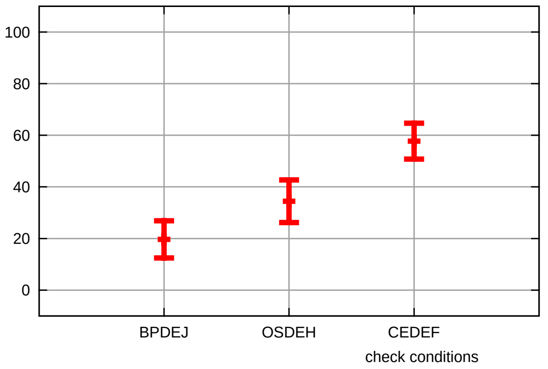
NRK - DIA
Average and 95% Confidence Intervals



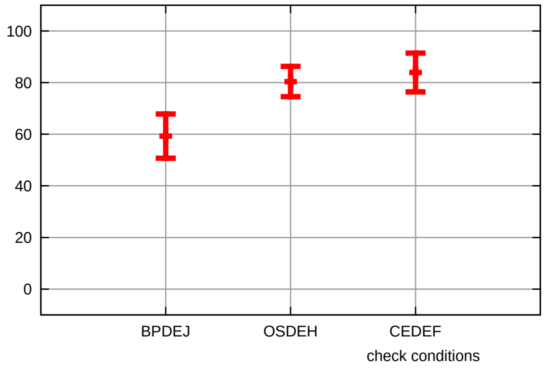
NRK - HOC
Average and 95% Confidence Intervals



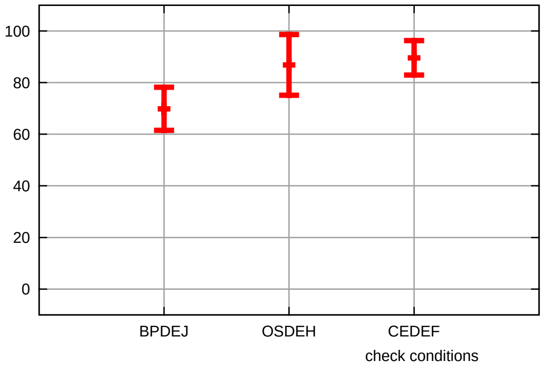
NRK - HRP
Average and 95% Confidence Intervals



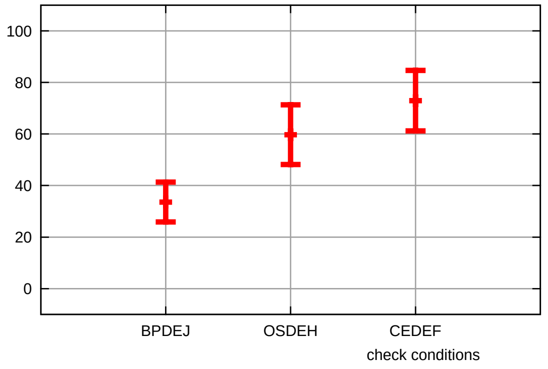
NRK - ORC
Average and 95% Confidence Intervals



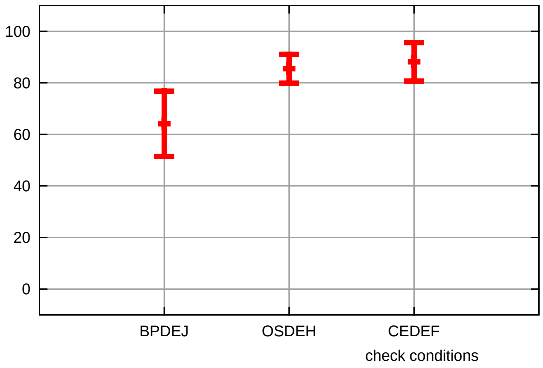
NRK - REA
Average and 95% Confidence Intervals



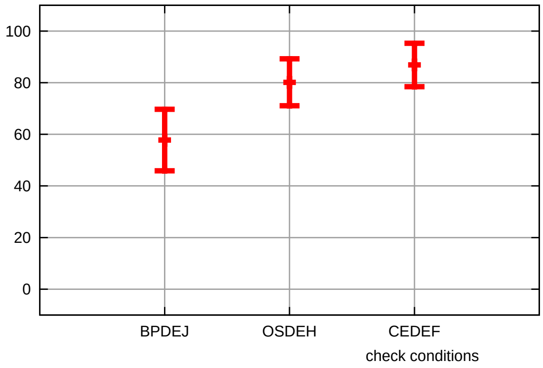
NRK - VEGA
Average and 95% Confidence Intervals



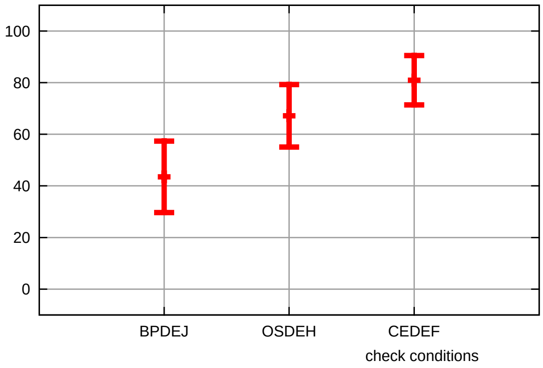
TVP - ACC
Average and 95% Confidence Intervals



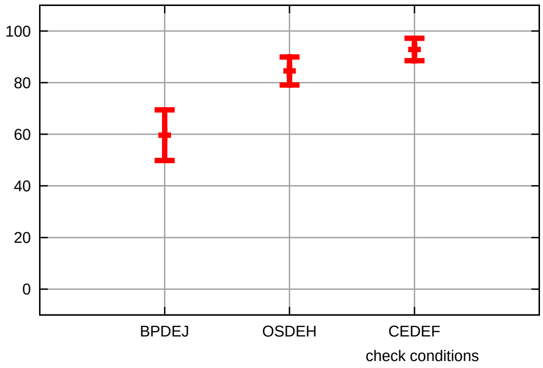
TVP - CAS
Average and 95% Confidence Intervals



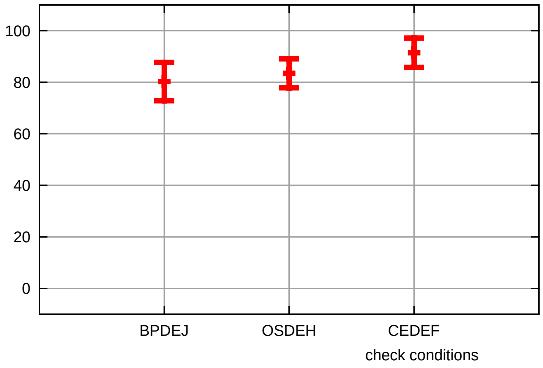
TVP - CLA
Average and 95% Confidence Intervals



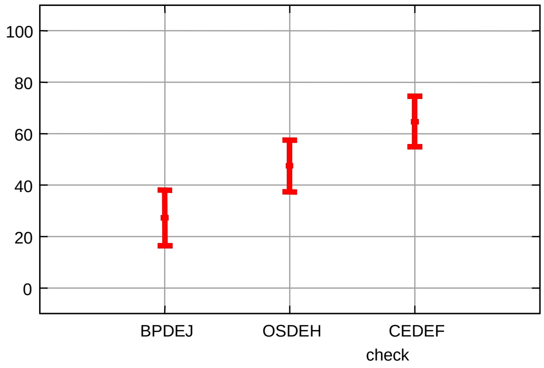
TVP - DIA
Average and 95% Confidence Intervals



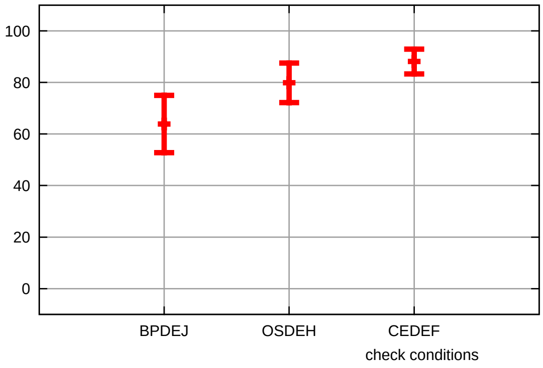
TVP - HOC
Average and 95% Confidence Intervals



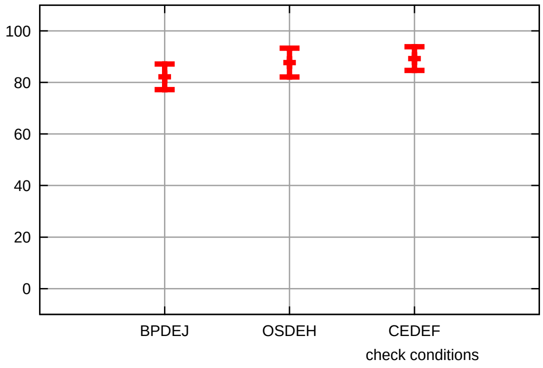
TVP - HRP
Average and 95% Confidence



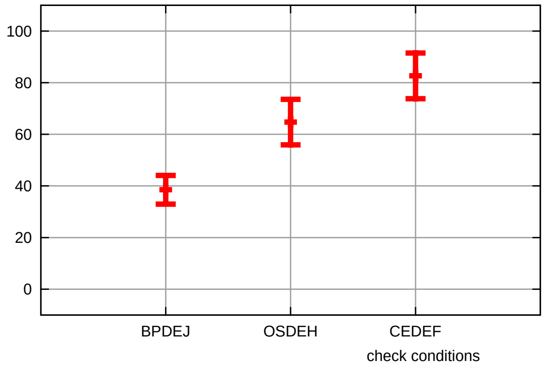
TVP - ORC
Average and 95% Confidence Intervals



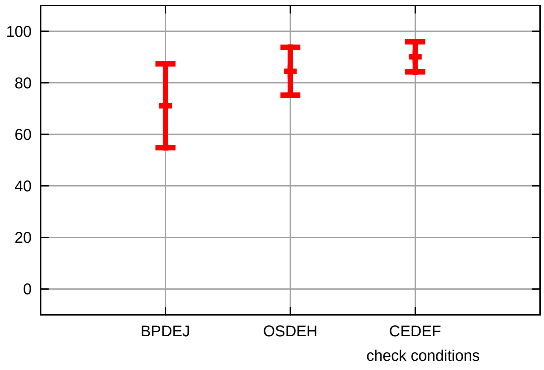
TVP - REA
Average and 95% Confidence Intervals



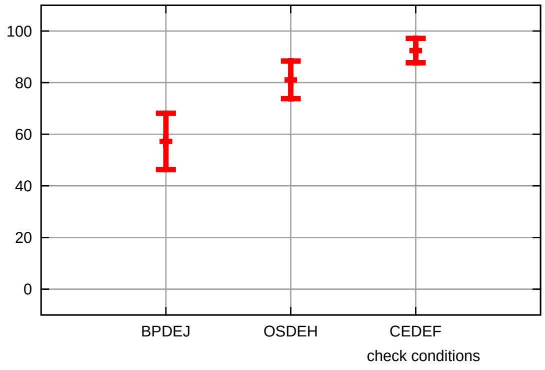
TVP - VEGA
Average and 95% Confidence Intervals



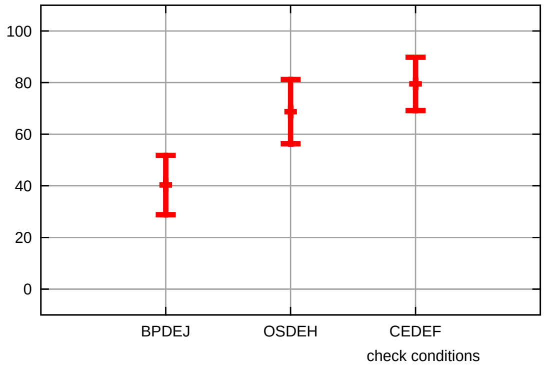
students - ACC
Average and 95% Confidence Intervals



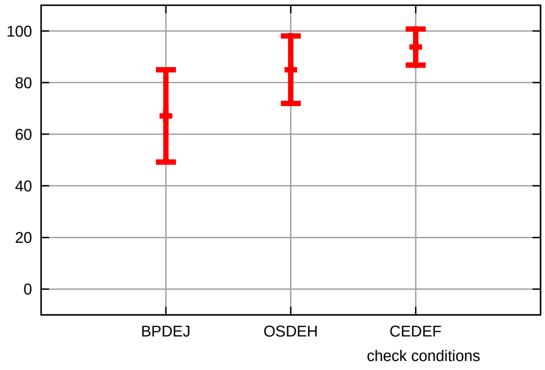
students - CAS
Average and 95% Confidence Intervals



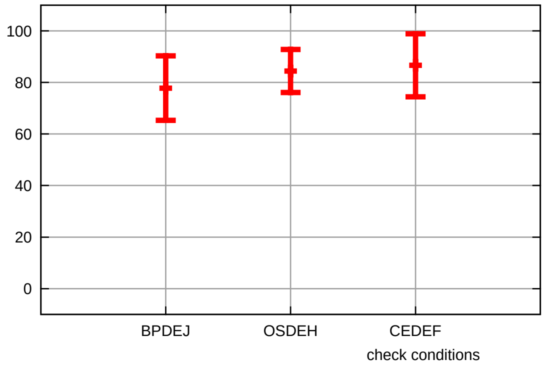
students - CLA
Average and 95% Confidence Intervals



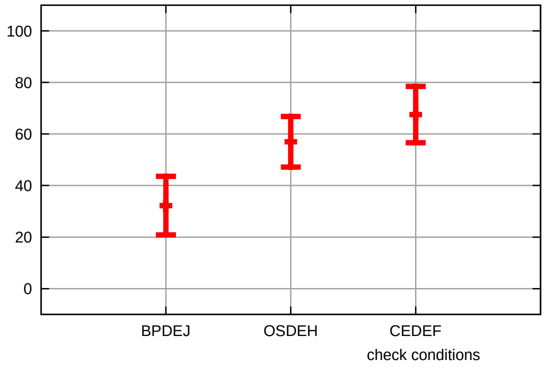
students - DIA
Average and 95% Confidence Intervals

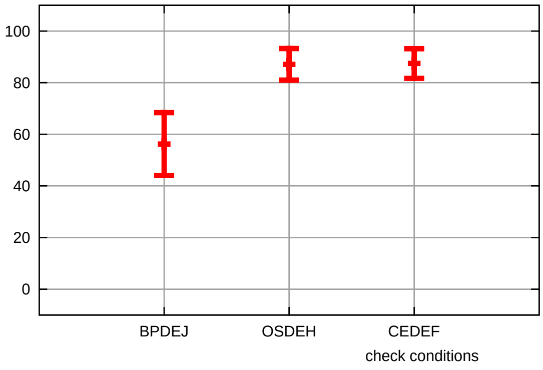


students - HOC
Average and 95% Confidence Intervals

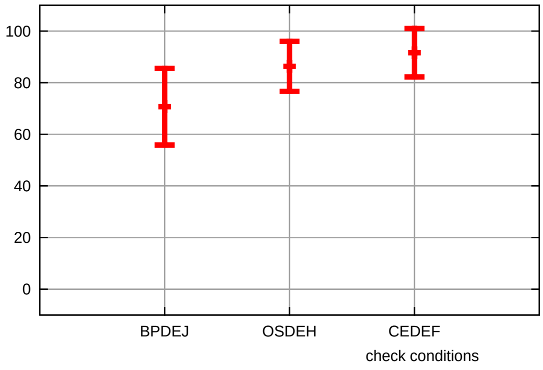


students - HRP
Average and 95% Confidence Intervals

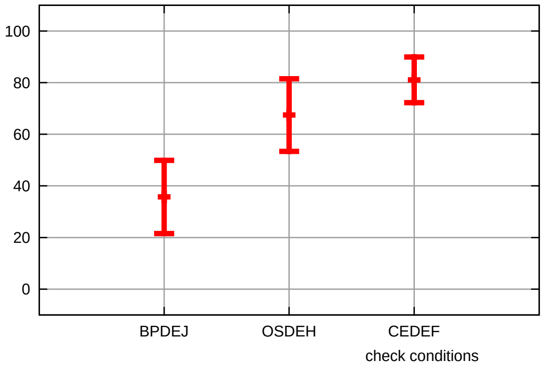




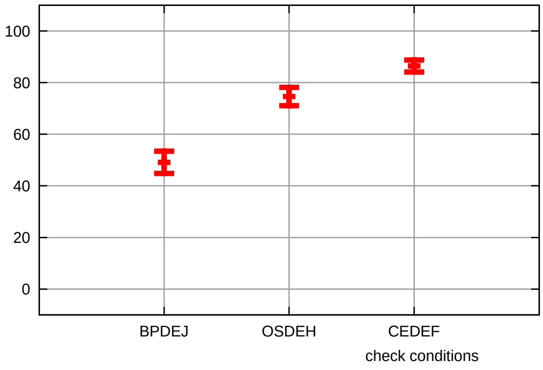
students - ORC
Average and 95% Confidence Intervals



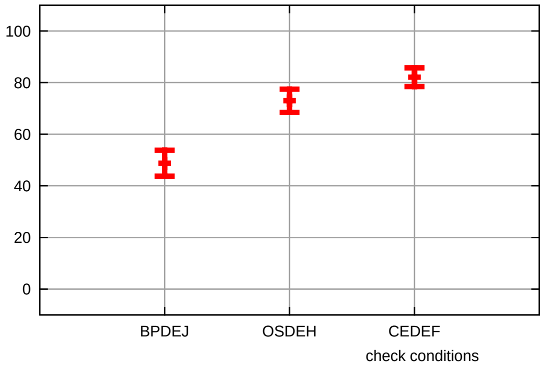
students - REA
Average and 95% Confidence Intervals



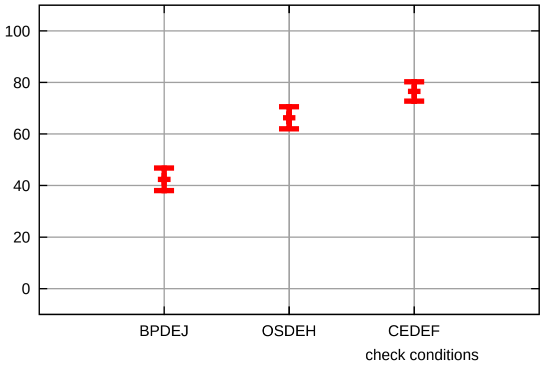
students - VEGA
Average and 95% Confidence Intervals



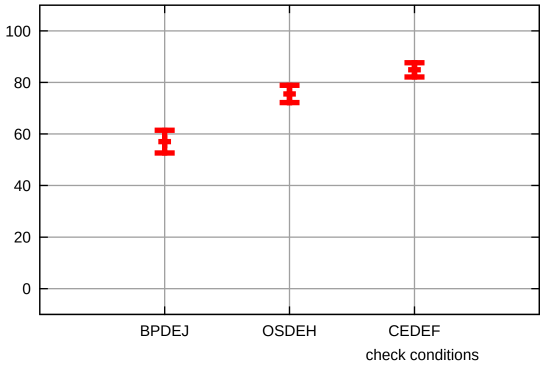
BBC - All Items
Average and 95% Confidence Intervals



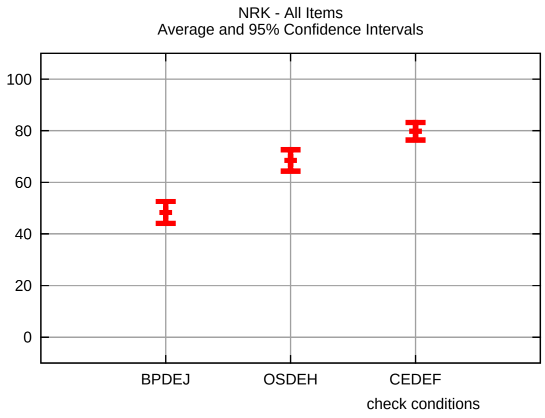
IRT_NRK_TDA - All Items
Average and 95% Confidence Intervals

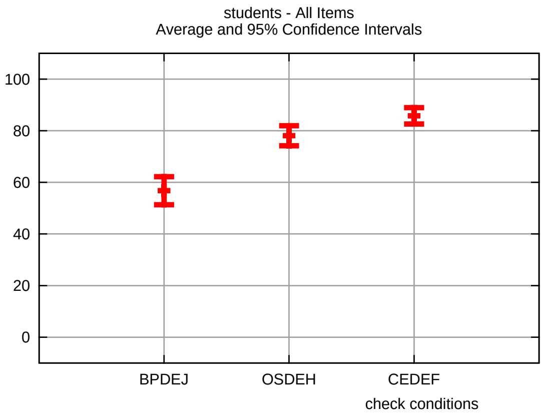


IRT - All Items
Average and 95% Confidence Intervals

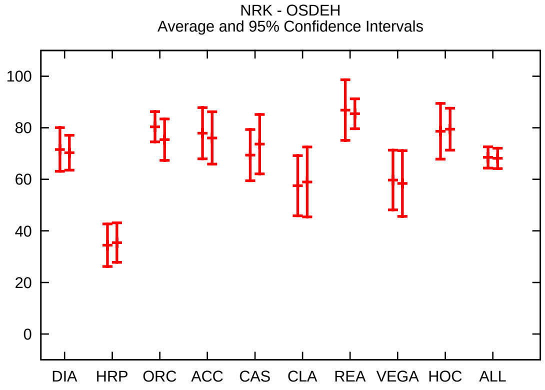


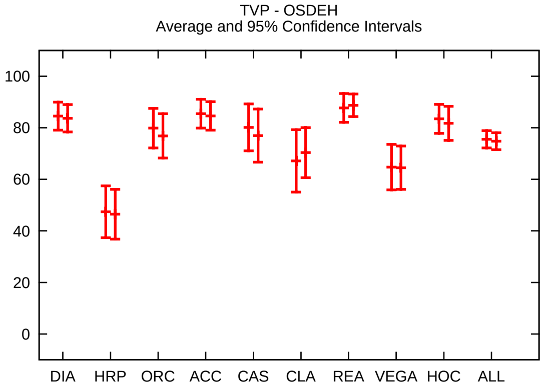
TVP - All Items
Average and 95% Confidence Intervals

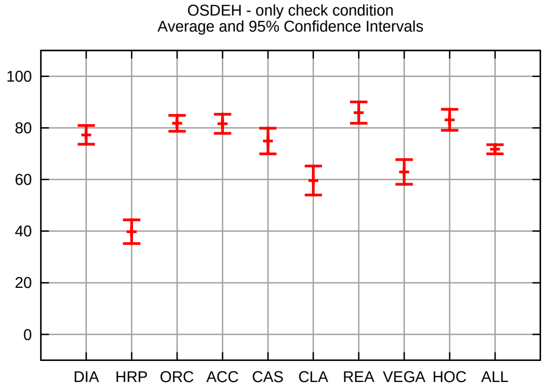




The following graphs show how the OSDEH cascade, which was used twice in NRK and TVP's test sequences compared against itself. The graph showing how OSDEH performed overall is also shown.







- - - - - End of Report - - - - - -