Joint TSG-S4#10 - SMG11#15 Meeting
February 28-March 3, 2000, Helsinki, Finland

Tdoc S4/SMG11 (00)0016~~98~~

Source: **Editor**[1]

Title: **~~Selection~~ Qualification Rules for AMR-WB (WB-5a version 0.3~~21~~)**

Document for: **Discussion**

Agenda Item:

This document contains a preliminary draft (skeleton) for the AMR-WB ~~selec~~qualification rules ~~to follow in the AMR-WB development process~~. It was prepared on the basis of the Selection Rules for the AMR-NB Selection Phase (AMR-NB Permanent Document 4b version 1.0, Tdoc SMG11 133/98). ~~assuming that the AMR-WB development would only involve a Selection Phase and no Qualification Phase. If a Pre-selection or Qualification Phase was also necessary, a separate Pre-Selection Rules document might be required.~~

A number of proposals and annexes already included in the AMR-NB Selection Rules were also included in this document to serve as a reference and a starting point~~preliminary proposal~~ for the final AMR-WB ~~selection~~ qualification rules. ~~The final rules will need to be updated according to the final content of the AMR-WB Selection Test Plan.~~

The following section lists the set of rules. The ~~selection~~ qualification procedure is described in section 2.

1. ~~Selection~~Qualification Rules:

Three basic rules are defined for the ~~selection~~qualification phase. The first two rules are eliminating rules intended to exclude all candidates failing to demonstrate full compliance with the AMR-WB Design Constraints defined in WB-4~~3~~ or presenting test results too far below the expected performance level. The third rule is not exactly a rule but a primary selection of Figures of Merit according to which the candidate performances will be compared as part of the ~~selection~~qualification test results analysis. These multiple criteria are intended to provide a more complete picture of the relative performances of the proposed solutions.

Each rule is further described in the following sections:

~~Selection~~Qualification Rule 1:

Any candidate not compliant with all Design Constraints defined in WB-4 (latest version approved by TSG-S4/SMG11) will be excluded.

~~Selection~~ Qualification Rule 2a:

Any candidate failing 50% or more of the test conditions contained in any of the following test sets will be excluded. A test is failed if the codec performance (measured MOS score or PoW) does not meet the requirement specification at the 95% confidence level.

List of test sets for Rule 2a:

Set #1:	all conditions (39)~~To be defined~~
Set #2:	all clean conditions (13)~~To be defined~~
Set #3:	all background noise conditions (26)~~To be defined~~
Set #4:	all conditions of application A (15)~~To be defined~~
Set #5:	all conditions of application B (12)~~To be defined~~
Set #6:	all conditions of application C, D, E (12)~~To be defined~~

The 50% threshold should be computed for each test set across the conditions tested by all listening laboratories performing an experiment included in this test set.

---

1 **Alain Ohana**
GSM North America Alliance
Mailing Address: PO Box 868075, Plano, TX 75086-8075, USA
Tel: 1 972 517 0709
Fax: 1 972 517 0709
Email: alain.ohana@pcs.bls.com

Selection Rule 2b:

~~Any candidate failing any single test condition in any experiment by more than 6 dBq will be excluded.~~

Any candidate failing severly in more than 10% of the test conditions contained in any of the following test sets will be excluded.

List of test sets for Rule 2b:
Set #1: all conditions (39)
Set #2: all clean conditions (13)
Set #3: all background noise conditions (26)
Set #4: all conditions of application A (15)
Set #5: all conditions of application B (12)
Set #6: all conditions of application C, D, E (12)

The 10% threshold should be computed for each test set across the conditions tested by all listening laboratories performing an experiment included in this test set.

A severe failure is defined by more than 6dBq MOS score difference or ΔPoW > 15% if applicable.

This criteria will only apply if either the equivalent Q value of the codec under test or the equivalent Q value of the reference codec is in the linear region of the MNRU curve and the~~o~~ test results in~~with~~:

- ΔMOS<-0.5 for any ACR test
- ~~-~~ ΔMOS <-1 for any DCR test based on MOS
- ΔMOS <-1 as well as ΔPoW > 15% for any DCR test based on PoW
- (ΔMOS = Codec MOS - Reference Codec MOS)

In addition, if the equivalent Q value of the reference codec is in the high Q saturation region, the 6 dBq threshold must be computed against the lower limit of the high Q saturation region. Likewise, if the equivalent Q value of the codec under test is in the low Q saturation region, the 6 dBq threshold must be computed against the higher limit of the low Q saturation region.

The saturation regions must be computed as the area of the MNRU curve where the slope is lower than 0.15/3=0.05 (corresponding to the points of the curve where a 3dBq step results in a MOS increase lower than 0.15).

Finally the rule will only apply if the test is consistently failed in the previously defined condition (6dBq and more than 0.5 or 1 ΔMOS) for a majority (>50%) of the listening laboratories performing the test (or failed for ~~2 labs out of 2,~~ 2 labs out of 3~~, or 3 labs out of 4~~).

Examples for the application of rule 2b are provided in Annex A.

Definition of ΔPoW is given in Annex C.

~~Selection~~ Qualification Rule 3: Figures of Merit:

A number of Figures of Merit will be used to analyze and compare the performance of the candidates. Corresponding rankings will be prepared and provided for information only. None of the Figures of Merit listed below is intended to serve as single selection criteria. SQ is entitled to define ~~a~~ preferred criteria for the quality assessment of the candidates.

The candidates will be ranked according to the following metrics:

- Number of majority Failures (2 failures out of 3 tests)
- Weighted ~~and Unweighted~~ ΔMOS (ΔMOS = Codec MOS - Reference MOS)
- Weighted ~~and Unweighted y~~ΔdBq (ΔdBq = Codec dBq - Reference dBq)
- Unweighted ΔPoW percentages (for the relevant conditions)

(Note that for both the ΔMOS and ΔdBq, the MOS or dBq value of the reference must be used rather that the MOS or dBq value of the requirement, unless it is for the purpose of a Figure of Merit restricted to the test failures. In the latter case, the MOS or dBq value of the requirement must be used instead, except for the cases when the requirement is defined in terms of PoW).

The following table lists the selected Figure of Merits and any limitation identified for this criteria:

Metric	Ranking Provided	Limitations
Weighted ΔdBq ~~Unweighted ΔdBq~~	Per experiment and across all experiments Per lab and across labs Full set of test results (Preferred FoM) and restricted to the failed tests only (ΔdBq computed with reference to the requirement in this case)	Arbitrary rule for evaluation of the dBq in the saturation region of the MNRU curve. For the FoM restricted to the failures only, the smaller the number of failures, the larger the uncertainty of the metric
Weighted ΔMOS ~~Unweighted ΔMOS~~	Per Experiment and Per lab (cannot be computed across labs and experiments) ~~Total computed with and without Exp. 4a~~ Full set of test results and restricted to failed tests	~~Possible influence of Experiment 4a if compiled with other experiments because of the large expected gain compared to EFR~~ The MOS rating is a non-linear function that will distort the full picture The smaller the number of failures, the larger the uncertainty of the metric
Number of majority Failures (2 failures out of 3 tests) ~~Number of Failures~~	Per Experiment and across all experiments ~~Per lab and a~~Across labs	Fails to demonstrate degree of failure
~~-~~ Unweighted ΔPoW percentages (for the relevant conditions)	Per Experiment and across all relevant experiments	tbd

Table 1: List of criteria selected for the evaluation of the test results

For the computation of the ΔdBq, the following rule will apply. If either the dBq value of the codec under test or the dBq value of the reference codec are in either saturation region of the MNRU curve as defined for Rule 2b, then the ΔdBq will be computed by replacing the saturation region of the MNRU curve by a linear approximation with slope 0.05 originated at Qmax, or Qmin. (See examples provided in Annex B).

For the rankings using a weighted metric, each test condition will carry a weight as defined in the following table. The table also provides a summary of the relative magnitude of each major test condition (Clean Speech and Background Noise~~, and Dynamic~~). The different Balance Factors may be ~~are~~ used to even up the relative weights of the different conditions or applications ~~GSM Full Rate, GSM Half Rate, GSM EDGE and 3G (?) test conditions~~ for each experiment.

When computing a weighted Figure of Merit, the weights should be applied to each test result in each experiment. ~~When combining the results across all listening laboratories, a second balance factor will be applied to take into account how many times the experiment has been performed. The values of these factors should be 1 for all experiments performed twice 2/3 for all experiments performed three times and ½ for experiments performed four times.~~

The test conditions with weight 0 should not be taken into account in ~~either~~ the weighted ~~or unweighted~~ Figures of Merit.

Experiment 1: Performances in Clean Speech (ACR)

Conditions	Application	Requirement	Weight
No Errors	A	Bet. Than G.722 48k	1,0
13 dB C/I	A	G.722 48k	1,0
10 dB C/I	A	EFR Degradation	1,0
7 dB C/I	A	EFR Degradation	1,0
4 dB C/I	A	EFR Degradation	1,0
No Errors	B	G.722 56k	1,0
19 dB C/I	B	G.722 56k	1,0
16 dB C/I	B	G.722 48k	1,0
13 dB C/I	B	G.722 48k	1,0
No Errors	C/D/E	G.722 64k	1,0
0.5%,0.0%	E	G.722 56k	1,0
1.0%,0.1% UL	E	G.722 48k	1,0
1.0%, 0.1% DL	E	G.722 48k	1,0
1.0%, 0.1% UL	E	-	0,0
		Total:	13,0
		Exp. Balance Factor:	1,0

Experiment 2: Performances in Background Noise conditions (DCR)

		Experiment 2a / Car Noise		Experiment 2b / Street Noise	
Conditions	Application	Requirement	Weight	Requirement	Weight
No Errors	A	G.722 48k 10% PoW	1,0	G.722 48k 10% PoW	1,0
13 dB C/I	A	G.722 48k 10% PoW	1,0	G.722 48k 10% PoW	1,0
10 dB C/I	A	EFR Degradation	1,0	EFR Degradation	1,0
7 dB C/I	A	EFR Degradation	1,0	EFR Degradation	1,0
4 dB C/I	A	EFR Degradation	1,0	EFR Degradation	1,0
No Errors	B	G.722 56k 10% PoW	1,0	G.722 56k 10% PoW	1,0
19 dB C/I	B	G.722 48k 10% PoW	1,0	G.722 48k 10% PoW	1,0
16 dB C/I	B	G.722 48k 10% PoW	1,0	G.722 48k 10% PoW	1,0
13 dB C/I	B	G.722 48k 10% PoW	1,0	G.722 48k 10% PoW	1,0
No Errors	C/D/E	G.722 64k	1,0	G.722 64k	1,0
0.5%,0.0%	E	G.722 56k	1,0	G.722 56k	1,0
1.0%,0.1% UL	E	G.722 48k	1,0	G.722 48k	1,0
1.0%, 0.1% DL	E	G.722 48k	1,0	G.722 48k	1,0
1.0%, 0.1% UL	E	-	0,0	-	0,0
		Total:	13,0	Total:	13,0
		Exp. Balance Factor:	0,5	Exp. Balance Factor:	0,5

Total weight Clean speech:	13,0
Total weight Background Noise:	13,0
Total weight:	26,0
Total weight for Application A:	10
Total weight for Application B:	8
Total weight for Application E:	8

Table 2: Weighting table for the Qualification rule 3

Proposal for integrating the conditions for which the requirements are expressed in Quality Degradation compared to EFR or PoW:

The MOS or dBq value of the reference codec should be used to compute the ΔMOS of ΔdBq when the requirement is expressed in PoW.

The equivalent MOS or dBq value of the transposed reference should be used to compute the ΔMOS of ΔdBq when the requirement is expressed in quality degradation compared to EFR. For example, if the quality degradation measured for EFR between 13 dB C/I and 10 dB C/I corresponds to a 0.2 MOS, then the reference for the test condition 10 dB C/I should be set to 0.2 MOS below the score obtained the AMR-WB codec under test at 13 dB C/I.

2. ~~Selection~~ Procedure:

The selection procedure will consist of~~in~~ the following steps:

1. The ~~selection~~Qualification test results will be presented and analyzed while keeping secret the identity of the candidates. Each candidate will be informed of the code used for its own solution and its solution only. The Qualification~~selection~~ rules 2a, 2b and 3 defined in the previous section will be applied at this stage.
2. After the review and discussion of the test results (as specified for rule 3), TSG-S4/SMG11 will try to reach a consensus on a quality ranking of the candidates.

3. Each candidate will then present its solution and show the compliance with the design constraints. All candidates not compliant with all design constraints will be excluded according to the ~~selection~~ Qualification rule 1.
4. The test results obtained by each candidate will then be revealed.
5. A final discussion and review of the solution characteristics and test results will take place.
6. S4/SMG11 will then try to reach a consensus on a list of candidates to ~~be selected~~keep for the Selection Phase~~AMR standardization~~.

3. References:

[1]: AMR-WB3: Adaptive Multi-Rate Wideband Speech Codec; Performance Requirements
Last version: 1.2 in S4-00090

[2]: AMR-WB4: Adaptive Multi-Rate Wideband Speech Codec; Design Constraints
Last version: 1.0 in Tdoc. S4-00087

Annex A: Examples for applicability of Rule 2b

Example 1:

ACR test where:
Lower limit of the linear region: Qmin = 3.0
Higher limit of the linear region: Qmax = 32.0
Reference Codec MOS score: Yref = 4.3
Codec under test MOS score: Ytest = 3.75
Reference Codec equi. Q Value: Qref = 31.0 (<Qmax, >Qmin)
Codec under test equi. Q Value: Qtest = 23.1 (<Qmax, >Qmin)
relevant delta Q: -ΔdBq = 31.0 - 23.1 = 7.9 (>6 dBq)
delta MOS: -ΔMOS = 4.3 - 3.75 = 0.55 (>0.5)

Conclusion: both Qref and Qtest are in the linear region, the codec fails the requirement by more than 6 dBq and 0.5 MOS, rule 2b applies and ~~the codec is excluded.~~the codec is judged to have a severe failure for this condition.

Example 2:

ACR test where:
Lower limit of the linear region: Qmin = 3.0
Higher limit of the linear region: Qmax = 28.0
Reference Codec MOS score: Yref = 4.3
Codec under test MOS score: Ytest = 3.75
Reference Codec equi. Q Value: Qref = 31.0 (>Qmax)
Codec under test equi. Q Value: Qtest = 23.1 (>Qmax, >Qmin)
relevant delta Q: -ΔdBq = 28.0 - 23.1 = 4.9 (<6 dBq)
delta MOS: -ΔMOS = 4.3 - 3.75 = 0.55 (>0.5)

Conclusion: Qref is in the saturation region, Qtest is in the linear region, the codec fails the requirement by more than 0.5 MOS but is not at 6 dBq from the relevant Q reference (Qmax), rule 2b applies and ~~the codec is not excluded.~~ the codec is judged not to have a severe failure for this condition.

Example 3:

ACR test where:
Lower limit of the linear region: Qmin = 3.0
Higher limit of the linear region: Qmax = 28.0
Reference Codec MOS score: Yref = 4.55
Codec under test MOS score: Ytest = 4.0
Reference Codec equi. Q Value: Qref = 36.4 (>Qmax)
Codec under test equi. Q Value: Qtest = 30.3~~5~~ (>Qmax, >Qmin)
relevant delta Q: -ΔdBq = 36.4 - 30.3 = 6.1 (>6 dBq)
delta MOS: -ΔMOS = 4.55 - 4.0= 0,.55 (> 0.5)

Conclusion: Qref and Qtest are in the saturation region, rule 2b does not apply and the codec is judged not to have a severe failure for this condition ~~not excluded~~ (even if -ΔdBq > 6 dBq, and -ΔMOS > 0.5)

Example 4:

DCR test where:
Lower limit of the linear region: Qmin = 3.0
Higher limit of the linear region: Qmax = 32.0
Reference Codec MOS score: Yref = 4.6

Codec under test MOS score: Ytest = 3.5
Reference Codec equi. Q Value: Qref = 31.0 (<Qmax, >Qmin)
Codec under test equi. Q Value: Qtest = 22.5 (<Qmax, >Qmin)
Reference Codec PoW perc. PoWref = 1 %
Codec under test PoW perc. PoWtest= 17%
relevant delta Q: -ΔdBq = 31.0 – 22.5 = 8.5 (>6 dBq)
delta MOS: -Δ_{MOS} = 4.6 - 3.5 = 1.1 (>1.0)
Relevant ΔPoW value -ΔPoW = 17 –1 = 16 % (>15%)

Conclusion: both Qref and Qtest are in the linear region, the codec fails the requirement by more than 6 dBq , 1.0 MOS, and 15 % PoW, rule 2b applies and the codec is judged to have a severe failure for this condition.

Example of Computation of the saturation region of the MNRU curve

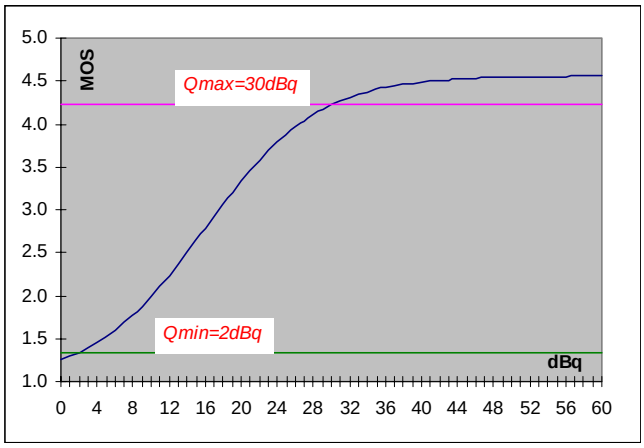
dBQ	MOS	Slope	Region
60	4.56	0.00050	Saturation
59	4.56	0.00058	Saturation
58	4.56	0.00068	Saturation
57	4.55	0.00080	Saturation
56	4.55	0.00094	Saturation
55	4.55	0.00110	Saturation
54	4.55	0.00129	Saturation
53	4.55	0.00152	Saturation
52	4.55	0.00178	Saturation
51	4.55	0.00208	Saturation
50	4.54	0.00244	Saturation
49	4.54	0.00286	Saturation
48	4.54	0.00335	Saturation
47	4.54	0.00393	Saturation
46	4.53	0.00460	Saturation
45	4.53	0.00538	Saturation
44	4.52	0.00629	Saturation
43	4.51	0.00736	Saturation
42	4.51	0.00859	Saturation
41	4.50	0.01003	Saturation
40	4.49	0.01170	Saturation
39	4.47	0.01363	Saturation
38	4.46	0.01586	Saturation
37	4.44	0.01842	Saturation
36	4.42	0.02137	Saturation
35	4.40	0.02474	Saturation
34	4.37	0.02858	Saturation
33	4.34	0.03294	Saturation
32	4.31	0.03784	Saturation
31	4.26	0.04332	Saturation
30	4.22	0.04940	Saturation
29	4.17	0.05609	Linear
28	4.11	0.06337	Linear
27	4.04	0.07118	Linear
26	3.96	0.07944	Linear
25	3.88	0.08803	Linear
24	3.79	0.09678	Linear
23	3.69	0.10548	Linear
22	3.58	0.11388	Linear
21	3.46	0.12169	Linear
20	3.33	0.12863	Linear
19	3.20	0.13440	Linear
18	3.07	0.13875	Linear
17	2.92	0.14146	Linear
16	2.78	0.14240	Linear
15	2.64	0.14153	Linear
14	2.50	0.13889	Linear
13	2.36	0.13460	Linear
12	2.23	0.12889	Linear
11	2.11	0.12199	Linear
10	1.99	0.11421	Linear
9	1.88	0.10583	Linear
8	1.78	0.09713	Linear
7	1.68	0.08838	Linear
6	1.60	0.07978	Linear
5	1.52	0.07150	Linear
4	1.46	0.06367	Linear
3	1.40	0.05637	Linear
2	1.34	0.04966	Saturation
1	1.30	0.04355	Saturation
0	1.26	0.03804	Saturation

The table besides presents for a typical experiment the correspondence between MOS scores and equivalent dBq.

The third column provides the value of the slope for each dBq value.

The last column identifies the saturation areas as the dBq values for which the slope is lower than 0.15 / 3 = 0.05.

The following figure shows the position of the saturation regions thresholds (Qmin and Qmax).

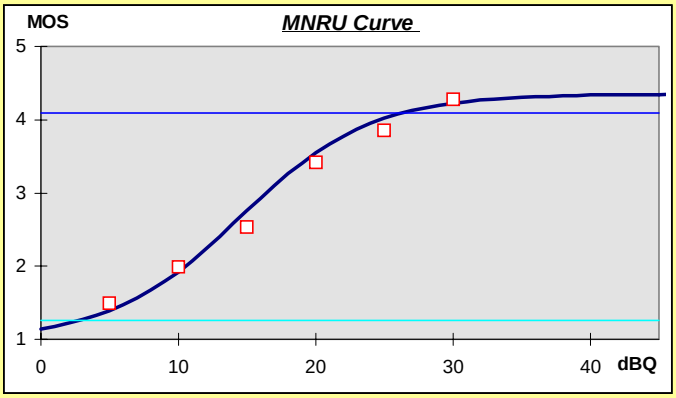


Annex B: Example of computation of the ΔdBq for rule 3

The following two examples are based on an hypothetical ACR for which the key parameters of the MNRU curve are reported in the figure below:

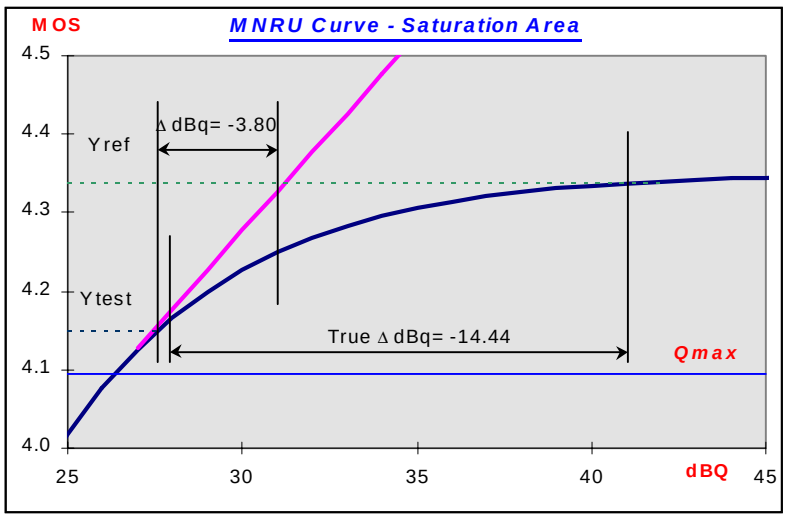
Ymax	**4.35**
S	0.05
M	14.55
Qmin	2.77
Qmax	26.34

MNRU	Qi	Yi	eqQ	dydQ
1	**30.00**	**4.28**	32.74	0.025
2	**25.00**	**3.85**	22.78	0.063
3	**20.00**	**3.42**	19.08	0.129
4	**15.00**	**2.53**	13.73	0.177
5	**10.00**	**1.99**	10.45	0.142
6	**5.00**	**1.49**	6.21	0.073



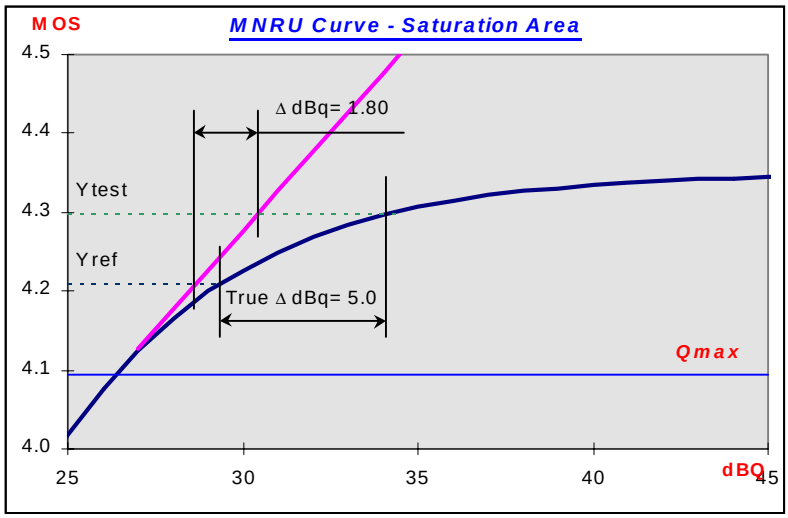
Example 1: The MOS and Q Value of the reference codec and codec under test are reported below and in the following figure:

Reference Codec MOS score: Yref = 4.34
Codec under test MOS score: Ytest = 4.15
Comparison based on MNRU Curve:
Reference Codec equi. Q Value: Qref = 42.03 (>Qmax)
Codec under test equi. Q Value: Qtest = 27.59 (>Qmax)
delta Q Value: ΔdBq = 27.59 ñ 42.03 = -14.44
Comparison based on Linearized saturation:
Reference Codec equi. Q Value: Qref = 31.26 (>Qmax)
Codec under test equi. Q Value: Qtest = 27.46 (>Qmax)
Relevant delta Q Value for Rule 3: ΔdBq = 27.46 ñ 31.26 **= -3.80**



Example 2: The MOS and Q Value of the reference codec and codec under test are reported below and in the following figure:

Reference Codec MOS score:	Yref = 4.21
Codec under test MOS score:	Ytest = 4.30
Comparison based on MNRU Curve:	
Reference Codec equi. Q Value:	Qref = 29.36 (>Qmax)
Codec under test equi. Q Value:	Qtest = 34.36 (>Qmax)
delta Q Value:	ΔdBq = 34.36 ñ 29.36 = 5.00
Comparison based on Linearized saturation:	
Reference Codec equi. Q Value:	Qref = 28.66 (>Qmax)
Codec under test equi. Q Value:	Qtest = 30.46 (>Qmax)
Relevant delta Q Value for Rule 3:	ΔdBq = 30.46 ñ 28.66 **= 1.80**



Annex C: Definition of ΔPoW

The „PoW"-votes are the votes where listeners rated the signal as MOS=2 („poor") and MOS=1 („bad").

In each specific condition under consideration, we characterise the result of the listening test in the following way:

- In a given condition, r % of all votes (i.e. a number of R votes out of a total of N votes; $r = 100 * R/N$) have rated the reference codec (Ref) being „poor" or „bad".
- In the same condition, c % of all votes (i.e. a number of C votes out of a total of N votes; $c = 100 * C/N$) have rated the codec under test (CuT) being „poor" or „bad".

Δ PoW is intended to be a relative measure for the number of votes that did prefer the Reference codec to the CuT.

Definition: In the situation described above, ΔPoW is defined to be

$\Delta PoW = (c - r)\ \%$, if $c - r > 0$, $\Delta PoW = 0\ \%$ otherwise

Application of this definition

Ideally for a good CuT, $c - r$ should be as small as possible and can be even negative. However, the PoW criterion is passed for $c<r$ by definition. Therefore, only the positive values of ΔPoW should be used, as stated in this definition.

The smaller the resulting figure of merit, the better is the performance of the codec with respect to it's reference codec (as used in each of the conditions considered for the calculation of the figure of merit, respectively).

End of Document