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

ISO/IEC JTC1/SC29/WG11
MPEG98/N2424

October 1998

Source: **MPEG Audio and Test subgroups**

Authors: **Pasi Ojala (Nokia Research Center), Henri Toukomaa (Nokia Research Center), Takehiro Moriya (NTT) and Oliver Kunz (FhG)**

Title: **Report on the MPEG-4 speech codec verification tests**

Status: **Approved**

1 Content

1 Content 1
2 Introduction 2
3 Test Format 2
4 Test method 3
4.1 Nokia site 3
4.2 NTT site 4
4.3 FhG site 4
5 Test excerpts 4
6 MPEG-4 codecs in test 4
6.1 Parametric coder 4
6.2 CELP speech coders 5
6.2.1 Narrowband CELP 5
6.2.1.1 CELP Mode VIII multirate 5
6.2.1.2 CELP Mode VIII scalable 5
6.2.2 Wideband CELP 5
6.2.2.1 Wideband CELP (fixed bit rate mode III) 5
6.2.2.2 CELP Bandwidth Scaleable mode 5
6.2.2.3 Optimised VQ + MPE 6
6.2.2.4 Optimised VQ + RPE 6
7 Reference codecs in test 6
7.1 FS1016 6
7.2 G.723.1 6
7.3 G.729 6
7.4 GSM EFR 6
7.5 G.722 6
7.6 MPEG-2 layer 3 7
8 Analysis of the test 7
8.1 Remarks on each experiment 7
8.1.1 Experiment 1 7
8.1.2 Experiment 2 7
8.1.3 Experiment 3 7
8.2 Differences between codecs 7

9 Test results 9
9.1 Overall performance 9
9.1.1 Nokia site 9
9.1.2 FhG site 12
9.1.3 NTT site 14
9.2 Performance in different languages 15
9.2.1 Nokia site 15
9.3 Performance item by item 23
9.3.1 Nokia site 23
9.3.2 FhG site 46
9.3.3 NTT site 57
10 References 69
11 Appendix A 70

2 Introduction

The MPEG-4 Audio coding tools cover a bit rate range from 2 kbit/s to 64 kbit/s with a corresponding subjective audio quality. Therefore, the MPEG-4 verification tests were carried out in several parts. The tests were related first of all Internet audio applications applying codecs with bit-rates ranging from 20 to 56 kbit/s, digital audio broadcasting on AM modulated bands with bit-rates of 16 to 24 kbit/s and speech applications. This document presents the MPEG-4 audio verification test results on speech coders. The performance of speech coders is evaluated in comparison with other standard coders. In this document the results of three independent test sites are presented.

3 Test Format

The test was defined in the Tokyo and Dublin meetings [1]. The following decisions were taken:

Due to the different technology and different band-width applied in the speech coders, the test had to be divided in three groups:

Test 1 contains narrow band parametric speech coders with 2 and 4 kbit/s. FS1016 was selected as a reference coder.

Codec	Bit rate (kbit/s)
Parametric	2, 4
Ref. FS1016	4.8

Table 1. Outline of test 1.

Test 2 contains narrow band CELP (NB-CELP) coders bit-rates ranging from 6 to 12 kbit/s. The test contains fixed bit-rate as well as bit-rate scalable coders. G.723.1, G.729 and GSM EFR coders operate as reference coders.

Codec	Bit rate (kbit/s)
CELP (Mode VIII multi rate)	6, 8.3, 12
CELP (Mode VIII scaleable)	8, 12
Ref. ITU-T G723.1	6.3
Ref. ITU-T G729	8
Ref. GSM-EFR	12.2

Table 2. Outline of test 2.

Test 3 contains the wide-band CELP (WB-CELP) coders with bit-rates ranging from 17.9 to 18.2 kbit/s as well as bandwidth scaleable CELP at 16 kbit/s. G.722 and MPEG2 layer 3 coders operate as reference coders.

Codec	Bit rate (kbit/s)
CELP (fixed rate Mode III)	18.2
CELP (BWscalable)	16
Optimized VQ+MPE	17.9
Optimized VQ +RPE	18.1
Ref. G.722	48, 56
Ref. MPEG-2 Layer 3	24

Table 3. Outline of test 3.

4 Test method

Absolute Category Rating (ACR) method according to ITU-T Recommendation P.800 was used. A five-grade scale for scoring was used:

ACR scale	
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

Table 4. Absolute Category Rating used in the verification tests.

The test sites and the number of valid listeners are shown below. Originally two more listeners (one for experiment 2 and one for experiment 3) took part in the test at FhG site. When analysing the results, all scores of these listeners in the experiment were discarded, since there were missing rating scores that could not be recovered.

	Japanese items	European items	
Test site	NTT	FhG	NRC
Native language of listeners	Japanese	German	Finnish
Number of listeners Exp 1	16	18	16
Number of listeners Exp 2	16	17	16
Number of listeners Exp 3	16	16	16

Table 5. The number of listeners in the verification tests in each test site.

4.1 Nokia site

All listeners were native Finnish, but had some basic knowledge about all the tested languages. The listeners were non-experts, and hired for this purpose outside Nokia. 6 of the listeners were females and 10 males. Age distribution was ranging from 21 to 39. Moreover, the same subjects listened to all three tests. The test were conducted in the specified order of test 1, test 2 and test 3. Sennheiser 580 headphones were applied.

The listener were trained for the test by first explaining the purpose of the test. Then the test procedure was discussed. Five samples were used for training of the listeners before the actual test. Test items were randomised separately for each listener.

4.2 NTT site

All listeners were native Japanese. Twelve listeners were female and four males. Age distribution was ranging from 20 to 45. The same subjects listened to all three tests. The test were conducted in the specified order of test 1, test 2 and test 3. STAX lambda Nova headphones were applied.

The listeners were all naive non-experts and trained for the test by first explaining the purpose of the test. Then the test procedure was discussed. Five samples were used for training of the listeners before the actual test.

4.3 FhG site

In FhG test site all listeners were native Germans. The test were conducted in the specified order of test 1, test 2 and test 3. STAX lambda Nova headphones were applied.

The listeners were all naive non-experts and trained for the test by first explaining the purpose of the test. Then the test procedure was discussed. Five samples were used for training of the listeners before the actual test.

5 Test excerpts

The MPEG 4 speech codec verification test was conducted with European and Japanese languages. In European test the languages were English, German and Swedish. A test panel selected the excerpts for the test from the NADIB speech sample database. Appendix A describes the selected Japanese and European items used in the tests. After the selection the samples were coded with the tested coders and MNRU noise reference samples were processed. Table 6 gives the number of test excerpts and reference MNRU items and the training items. In addition, the number of tested codecs is listed. Listening test specification document describes the responsibilities of the process [1].

	Parametric	NB-CELP	WB-CELP
CODEC	3	8	7
MNRU	4	4	4
Test excerpts	15	15	14
Training	5	5	5

Table 6. Number of test excerpts in each verification test.

6 MPEG-4 codecs in test

This section gives a short description of the MPEG 4 codecs tested in this verification test. The mode III and VIII, which are referred in this document, are as follows: Mode III is a combination of SQ and RPE, and mode VIII is a combination of VQ and MPE.

6.1 Parametric coder

MPEG-4 parametric speech coder uses Harmonic Vector eXcitation Coding (HVXC) algorithm where harmonic coding of LPC residual signals for voiced segments and Vector eXcitation Coding (VXC) for unvoiced segments are used. Pitch and speed change functionality are supported. The coder operates at 2.0 and 4.0 kbit/s of fixed bit rate mode and at less than 2.0 kbit/s of variable rate mode. 2.0 kbit/s decoding is possible using not only 2.0 kbit/s bitstream but also 4.0 kbit/s bitstream. The frame length is 20 ms, and one of four different algorithmic delays, 33.5 ms, 36ms, 53.5 ms,

56 ms can be selected. In the verification tests, 2.0 and 4.0 kbit/s fixed bit rates with 36 ms delay were used.

6.2 CELP speech coders

In the MPEG-4 CELP Audio decoder speech is generated by predicting the speech signal at the output using a Linear Prediction Filter. Its coefficients, which are extracted from the bitstream, are either quantised using Scalar Quantisation (SQ) or Vector Quantisation (VQ). The LPC filter is driven by an excitation module that can either be Regular Pulse Excitation (RPE) or Multi-Pulse Excitation (MPE). The bit rate can be selected in discrete steps or, when FineRate Control is enabled, any arbitrary bit rate can be generated within the range of the discrete steps.

6.2.1 Narrowband CELP

The Narrowband CELP coders use the combination of the MPE tool and the LSP-VQ tool and operate at a sampling rate of 8 kHz. Fine rate control is not being used in the test.

6.2.1.1 CELP Mode VIII multirate

Three fixed bit rates of 6.0, 8.3 and 12 kb/s, were used to evaluate the coding quality. The frame length was 20 ms for 6.0 and 8.3 kb/s and 10 ms for 12 kb/s. The length of look-ahead was 5 ms.

6.2.1.2 CELP Mode VIII scalable

The scalable operation of 6-kb/s core and three 2-kb/s enhancement layers was used. Two bit rates of 8.0 and 12.0 kb/s were evaluated. The frame length was 20 ms and the delay was 25 ms.

6.2.2 Wideband CELP

The Wideband CELP coders all operate at a sampling rate of 16 kHz enabling a bandwidth of 7.5 kHz.

6.2.2.1 Wideband CELP (fixed bit rate mode III)

Bit rate:	18200 bit/s
Quantisation mode:	Scalar Quantiser
Excitation:	RPE
FineRate Control:	off
Frame length:	15 ms
Delay:	18.75 ms

6.2.2.2 CELP Bandwidth Scaleable mode

Bit rate:	6000 bit/s (8 kHz part) +
	10000 bit/s (BWS extension)
Quantisation mode:	Vector Quantiser
Excitation:	MPE
FineRate Control:	off
Frame length:	20 ms
Delay:	30 ms

6.2.2.3 Optimised VQ + MPE

Bit rate:	17900 bit/s
Quantisation mode:	Vector Quantiser
Excitation:	MPE
FineRate Control:	off
Frame length:	20 ms
Delay:	25 ms

6.2.2.4 Optimised VQ + RPE

Bit rate:	18100 bit/s
Quantisation mode:	Vector Quantiser
Excitation:	RPE
FineRate Control:	on
Frame length:	15 ms
Delay:	33.75 ms

7 Reference codecs in test

This section gives a short description of the reference codecs utilised in this verification test.

7.1 FS1016

FS1016 coder is a US Federal Standard. It uses CELP algorithm operating at 4.8 kbit/s. Frame length is 30 ms and look-ahead length is 7.5 ms, resulting in the algorithmic delay of 37.5 ms.

7.2 G.723.1

G.723.1 was used as a reference in experiment 2. The G723.1 is a speech encoder recommended by ITU-T for multimedia communication at 5.3 and 6.3 kbit/s. In this test the 6.3 kbit/s version was used. This encoder was optimised for encoding speech signals with a high quality for a limited amount of complexity. The frame length is 30 ms with an additional look ahead of 7.5 ms, resulting in a total algorithmic delay of 37.5 ms.

7.3 G.729

G.729 was used as a reference in experiment 2. The G729 is a speech encoder recommended by ITU-T for multimedia communication at 8 kbit/s. This encoder was optimised for encoding speech signals with a high quality for a limited amount of complexity. The frame length is 10 ms with an additional look ahead of 5 ms, resulting in a total algorithmic delay of 15 ms.

7.4 GSM EFR

GSM EFR was used as a reference in experiment 2. The GSM EFR is a speech encoder recommended by ETSI. The codec operates as 12.2 kbit/s. The frame length is 20 ms without look ahead, resulting in an algorithmic delay of 20 ms.

7.5 G.722

G.722 coders were used as a reference in experiment 3. The G.722 is a generic audio coder recommended by ITU-T for multimedia communication. In this test, the

reference coder was operating at bitrates 48 and 56 kbit/s. Delay for the G.722 coder is 1.5 ms.

7.6 MPEG-2 layer 3

MPEG-2 layer 3 was used as a reference in experiment 3 operating at 24 kbit/s. The delay is 210 ms.

8 Analysis of the test

8.1 Remarks on each experiment

8.1.1 Experiment 1

MPEG-4 HVXC at both 2.0 and 4.0 kbit/s outperform the reference codec FS1016 at 4.8 kbit/s. Additionally, the HVXC coder has functionality, such as pitch and speed change and bit-rate scalability.

8.1.2 Experiment 2

The MPEG-4 NB-CELP coder with bit-rate ranging from 6 to 12 kbit/s provides competitive quality compared with the speech coding standards that were optimised for a single specific bit-rate.

Furthermore, the tested MPEG-4 CELP coder offers bit-rate scalability. The speech quality can be improved step-by-step by adding enhancement layers on top of the base layer coder.

There are some differences in quality depending on the tested language and input items.

8.1.3 Experiment 3

MPEG-4 WB-CELP coders for wide-band speech signals provide competitive quality compared with G.722 at 48 kbit/s and MPEG-2 Layer III at 24 kbit/s, as far as speech signals are concerned, at the bit-rate of 18 kbit/s with additional functionality, such as bit-rate, bandwidth and complexity scalability.

There are some differences in quality depending on the tested language and input items.

8.2 Differences between codecs

In Nokia site, the student tables for all three tests have been computed, together with the ranking of codecs based on the mean grades. From these data it is possible to build the NSSD (Next Statistically Significant Difference) matrixes for each case. Tables 7, 8 and 9 present the student tables and NSSD matrixes for tests 1, 2 and 3, respectively. The codecs having the average MOS score in the same column do not have any statistically significant difference in quality, where as the codecs having the score in different columns are significantly different. For example, in Table 7 the difference with MNRU 20 and FS1016 is not significant, while the difference between FS1016 and Parametric coder with 2.0 kbit/s is statistically significant.

SCORE

Tukey B[a]

Codec	N	Subset for alpha = .05					
		1	2	3	4	5	6
MNRU 10	240	1.20					
MNRU 20	240		2.05				
Ref. FS1016 4.8 kbit/s	240		2.19				
Parametric 2 kbit/s	240			2.53			
Parametric 4 kbit/s	240				2.92		
MNRU 30	240					3.20	
MNRU 40	240						4.20

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 240.000

Table 7. NSSD matrix for test 1 (Parametric)

SCORE

Tukey B[a]

Codec	N	Subset for alpha = .05						
		1	2	3	4	5	6	7
MNRU 10	240	1.13						
MNRU 20	240		1.80					
CELP (Mode VIII multi rate) 6 kbit/s	240			2.70				
CELP (Mode VIII scaleable) 8 kbit/s	240				2.98			
Ref. ITU-T G.723.1 6.3 kbit/s	240				2.99			
MNRU 30	240				3.00			
CELP (Mode VIII multi rate) 8.3 kbit/s	240					3.25		
CELP (Mode VIII scaleable) 12 kbit/s	240					3.38		
Ref. ITU-T G.729 8 kbit/s	240					3.38		
CELP (Mode VIII multi rate) 12 kbit/s	240						3.69	
Ref. GSM-EFR 12.2 kbit/s	240							3.91
MNRU 40	240							3.93

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 240.000

Table 8. NSSD matrix for test 2 (NB-CELP)

SCORE

Tukey B[a]

Codec	N	Subset for alpha = .05							
		1	2	3	4	5	6	7	8
MNRU 10	224	1.10							
MNRU 20	224		1.79						
MNRU 30	224			2.76					
CELP (fixed rate Mode III) 18 .2 kbit/s	224				3.14				
CELP (BW scalable) 16 kbit/s	224					3.41			
Ref. MPEG-2 Layer III 24 kbit/s	224					3.47			
Optimized VQ+RPE 18.1 kbit/s	224					3.53	3.53		
Ref. G.722 48 kbit/s	224					3.55	3.55		
Optimized VQ+MPE 17.9 kbit/s	224						3.74	3.74	
MNRU 40	224							3.92	3.92
Ref. G.722 56 kbit/s	224								4.02

Means for groups in homogeneous subsets are displayed.

a. Uses Harmonic Mean Sample Size = 224.000

Table 9. NSSD matrix for test 3 (WB-CELP)

It should be noted that in MPEG standards only the decoder is normative and that the MPEG-4 codecs supplied for these tests are developmental and further optimisation will continue.

9 Test results

In this section the test results are presented in detail. First the overall results are covered. The performance of the tested codecs was analysed in different European languages. Section 9.2 presents these results. Finally the item-by-item results are presented in section 9.3.

9.1 Overall performance

Means and confidence intervals for each of the codecs were computed, to evaluate their overall performance with all test items. The results of each test site are presented in separate sections. 9.1.1 contains the European language test in Nokia site. Section 9.1.2 cover the same test conducted in Fraunhofer site. And finally, section 9.1.3 presents the Japanese test conducted in NTT site.

9.1.1 Nokia site

The results of Parametric, NB-CELP and WB-CELP are shown in Tables 10, 11 and 12, and graphically in Figures 1, 2 and 3, respectively.

Descriptives

			N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
							Lower Bound	Upper Bound
SCORE	Codec	Parametric 2 kbit/s	240	2.53	.91	.059	2.41	2.64
		Parametric 4 kbit/s	240	2.92	1.02	.066	2.79	3.05
		Ref. FS1016 4.8 kbit/s	240	2.19	.83	.053	2.09	2.30
		MNRU 10	240	1.20	.44	.028	1.14	1.25
		MNRU 20	240	2.05	.76	.049	1.95	2.14
		MNRU 30	240	3.20	.90	.058	3.08	3.31
		MNRU 40	240	4.20	.79	.051	4.10	4.30
		Total	1680	2.61	1.21	.029	2.55	2.67

Table 10. Results of the listening test 1 (Parametric).

Descriptives

			N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
							Lower Bound	Upper Bound
SCORE	Codec	CELP (Mode VIII multi rate) 6 kbit/s	240	2.70	.78	.051	2.60	2.80
		CELP (Mode VIII multi rate) 8.3 kbit/s	240	3.25	.84	.054	3.14	3.36
		CELP (Mode VIII multi rate) 12 kbit/s	240	3.69	.87	.056	3.58	3.80
		CELP (Mode VIII scaleable) 8 kbit/s	240	2.98	.74	.048	2.88	3.07
		CELP (Mode VIII scaleable) 12 kbit/s	240	3.38	.82	.053	3.27	3.48
		Ref. ITU-T G.723.1 6.3 kbit/s	240	2.99	.78	.050	2.89	3.09
		Ref. ITU-T G.729 8 kbit/s	240	3.38	.81	.052	3.28	3.48
		Ref. GSM-EFR 12.2 kbit/s	240	3.91	.80	.051	3.81	4.01
		MNRU 10	240	1.13	.34	.022	1.08	1.17
		MNRU 20	240	1.80	.72	.047	1.71	1.89
		MNRU 30	240	3.00	.97	.063	2.87	3.12
		MNRU 40	240	3.93	.87	.056	3.82	4.04
		Total	2880	3.01	1.12	.021	2.97	3.05

Table 11. Results of the listening test 2 (NB-CELP).

Descriptives

			N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
							Lower Bound	Upper Bound
SCORE	Codec	CELP (fixed rate Mode III) 18 .2 kbit/s	224	3.14	1.00	.067	3.01	3.27
		CELP (BW scalable) 16 kbit/s	224	3.41	.93	.062	3.29	3.53
		Optimized VQ+MPE 17.9 kbit/s	224	3.74	.92	.061	3.62	3.86
		Optimized VQ+RPE 18.1 kbit/s	224	3.53	1.02	.068	3.39	3.66
		Ref. G.722 48 kbit/s	224	3.55	.96	.064	3.43	3.68
		Ref. G.722 56 kbit/s	224	4.02	.88	.058	3.91	4.14
		Ref. MPEG-2 Layer III 24 kbit/s	224	3.47	.97	.065	3.35	3.60
		MNRU 10	224	1.10	.32	.021	1.06	1.14
		MNRU 20	224	1.79	.71	.047	1.69	1.88
		MNRU 30	224	2.76	.80	.054	2.66	2.87
		MNRU 40	224	3.92	.97	.065	3.79	4.04
		Total	2464	3.13	1.24	.025	3.08	3.18

Table 12. Results of the listening test 3 (WB-CELP).

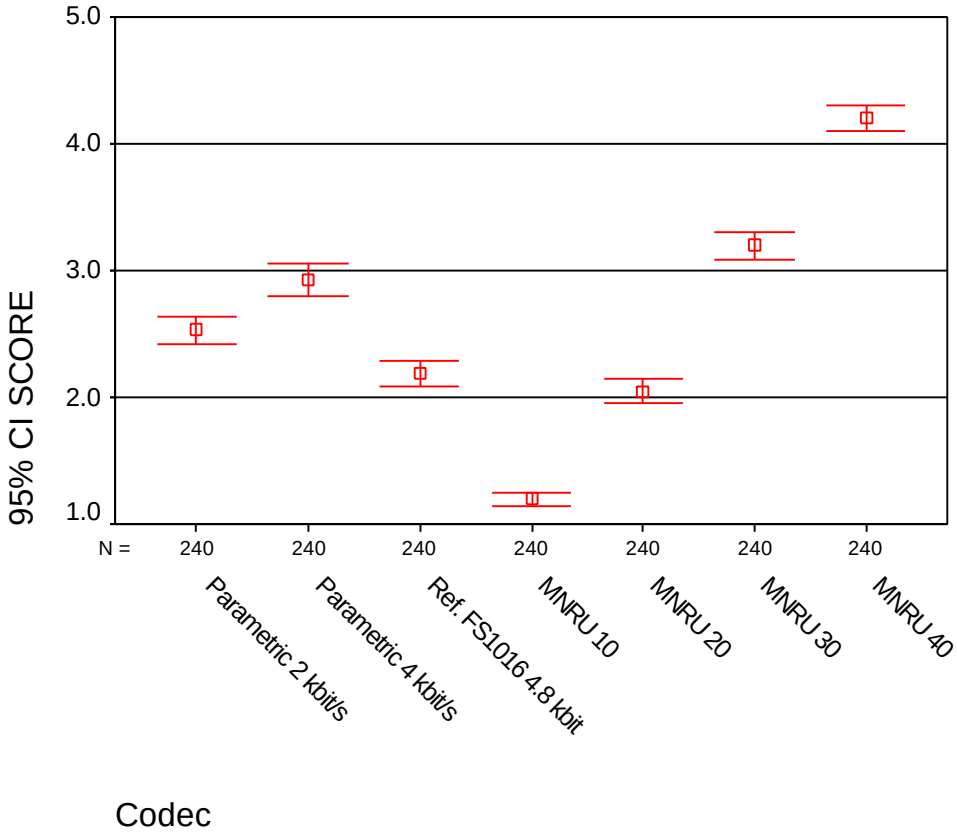


Figure 1. Overall results of the listening test 1 (Parametric).

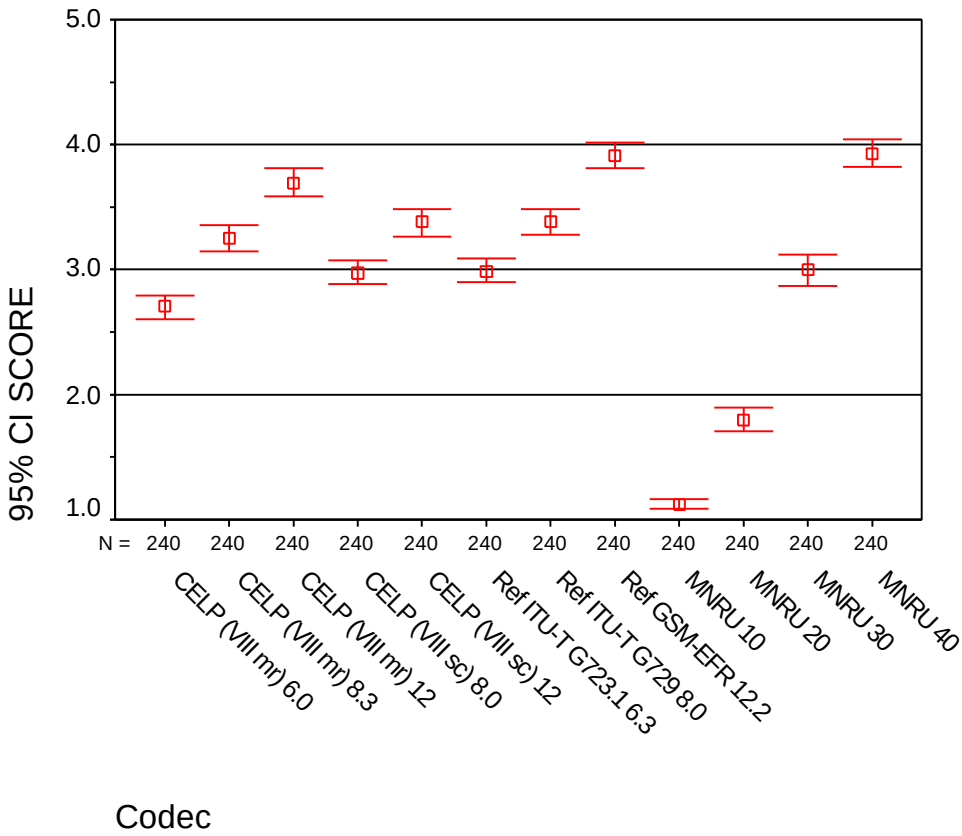


Figure 2. Overall results of the listening test 2 (NB-CELP).

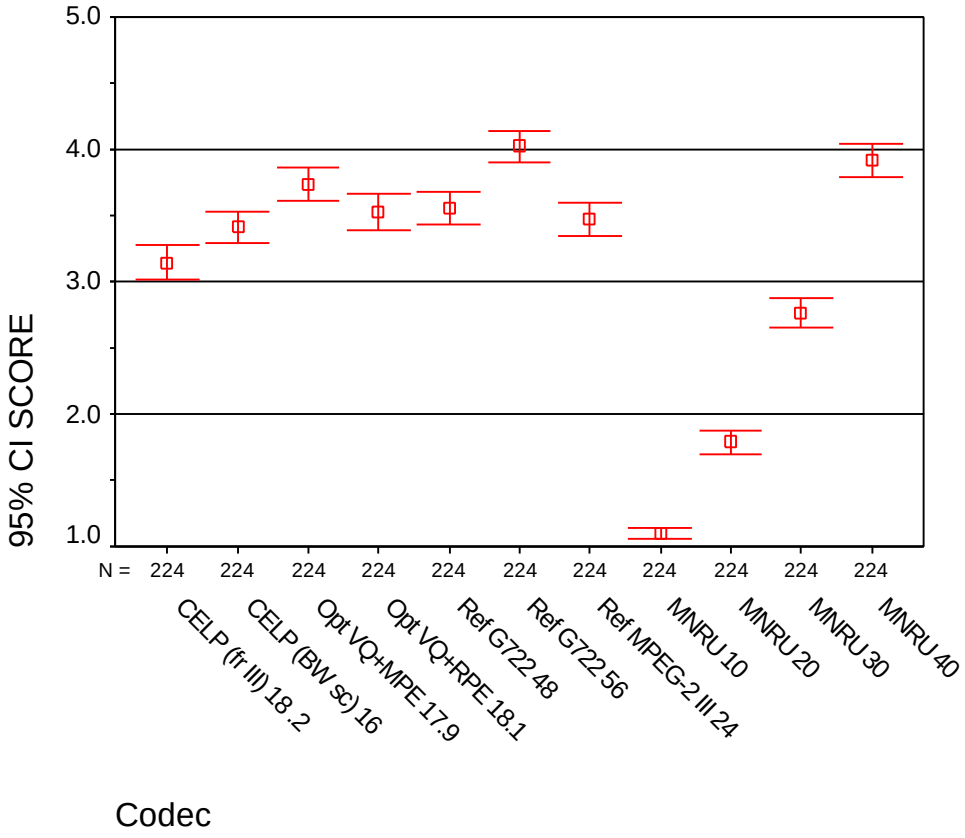


Figure 3. Overall results of the listening test 3 (WB-CELP).

9.1.2 FhG site

The results of Parametric, NB-CELP and WB-CELP are shown in Figures 4, 5 and 6, respectively. In Figure 5, CELP 8.0 kbit/s should be written as CELP 8.3 kbit/s.

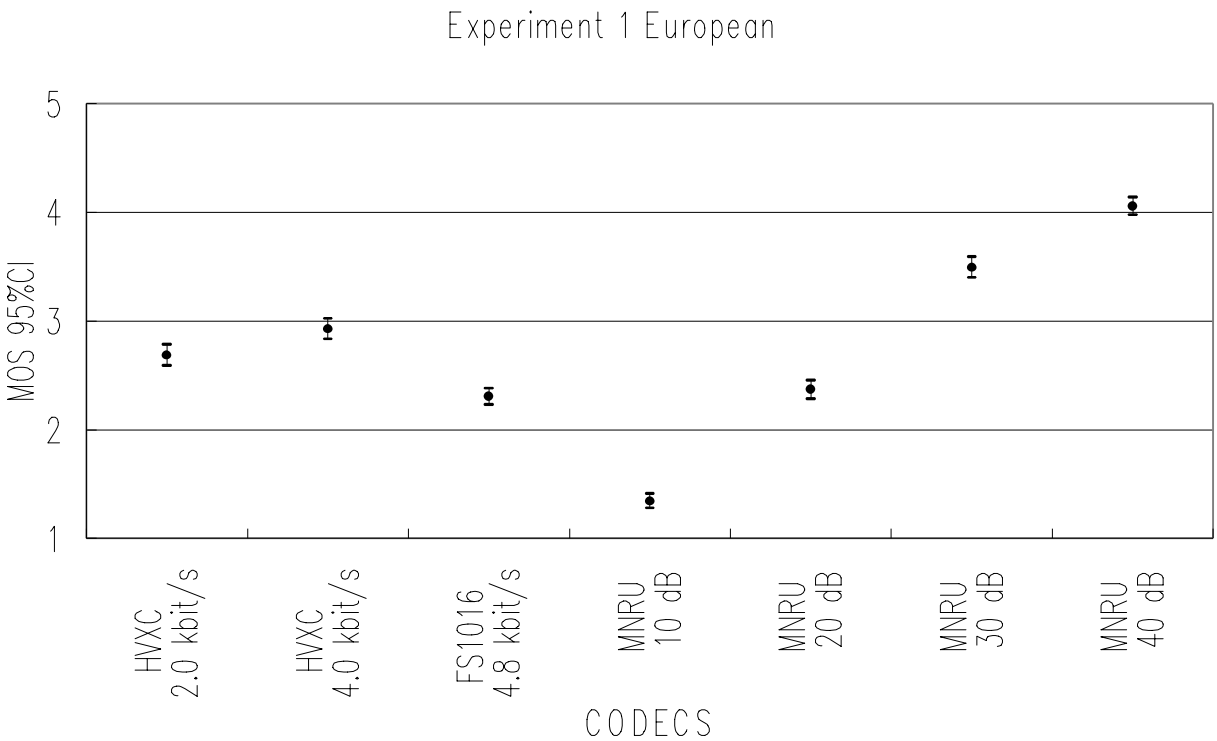


Fig. 4. MOS averaged for all European items.

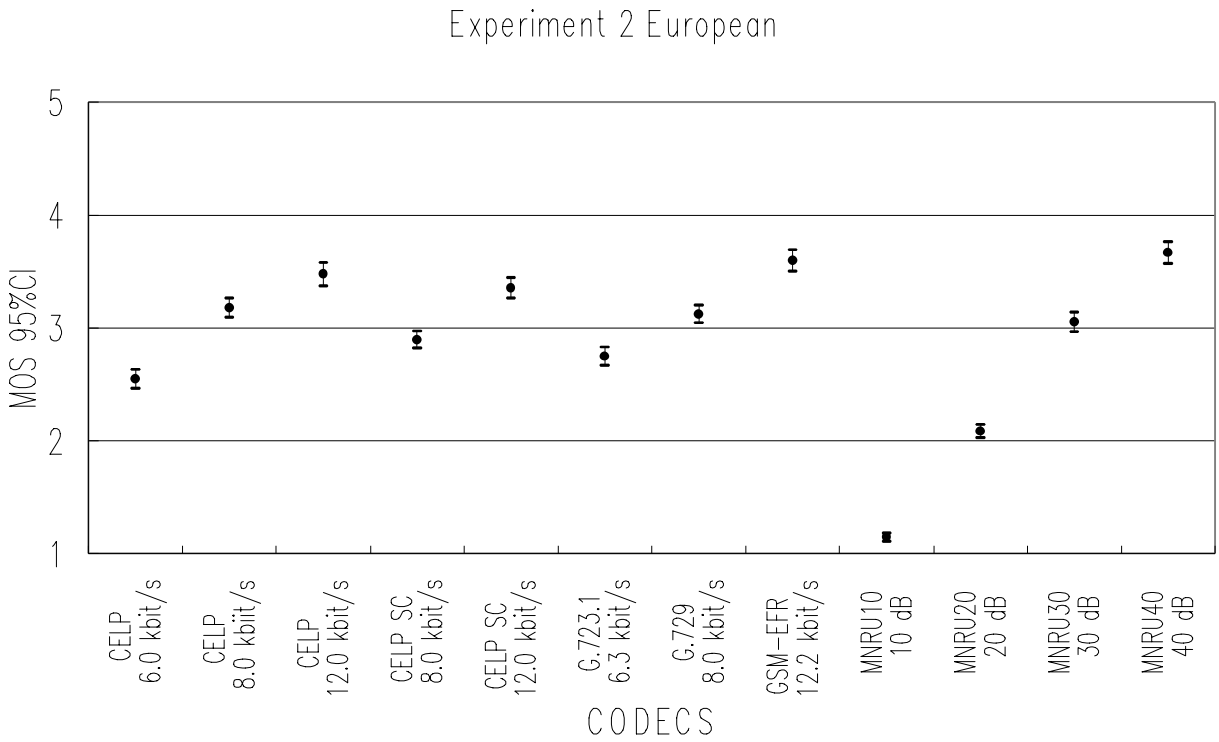


Fig. 5 MOS averaged for all European items.

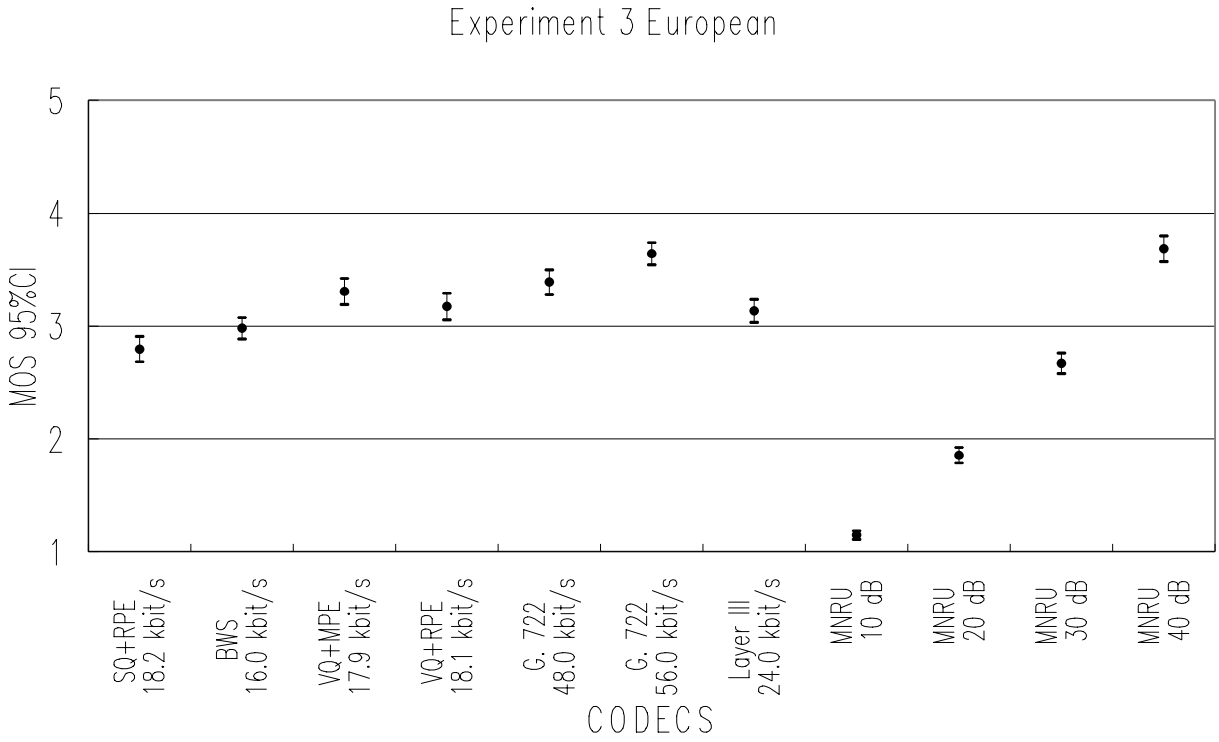


Fig. 6 MOS averaged for all European items.

9.1.3 NTT site

The results of Parametric, NB-CELP and WB-CELP are shown in Figures 7, 8 and 9, respectively.

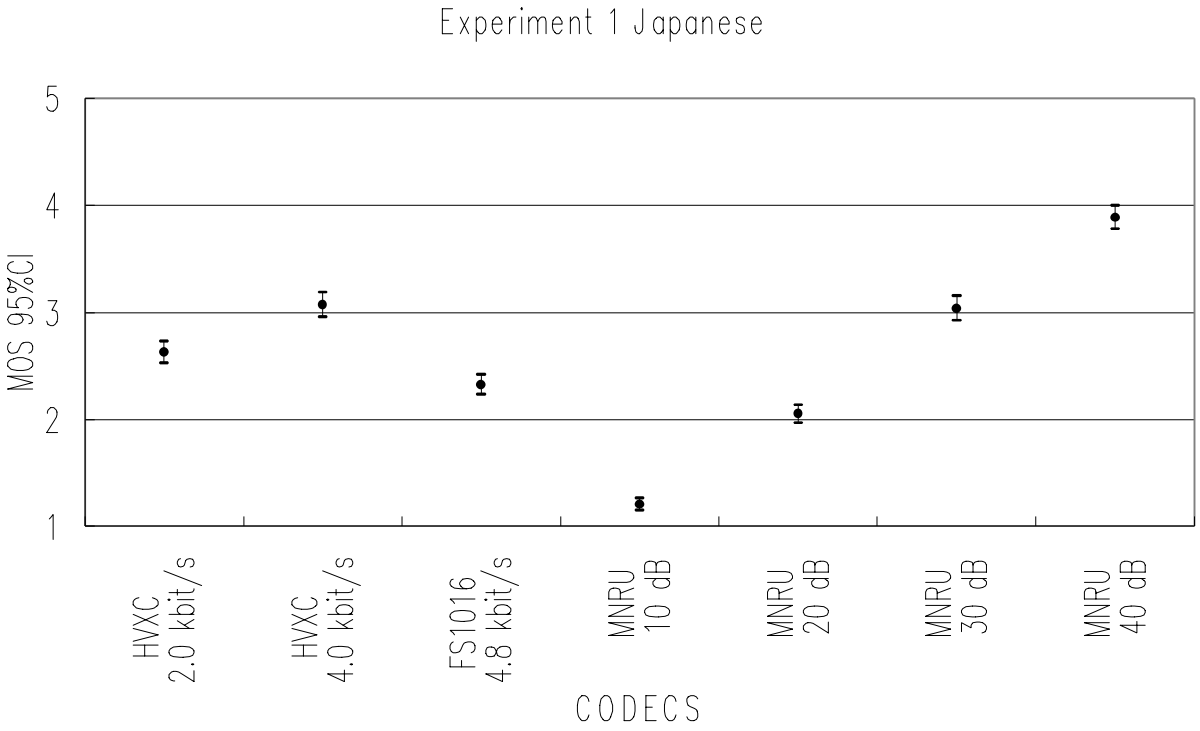


Fig. 7 MOS averaged for all Japanese items.

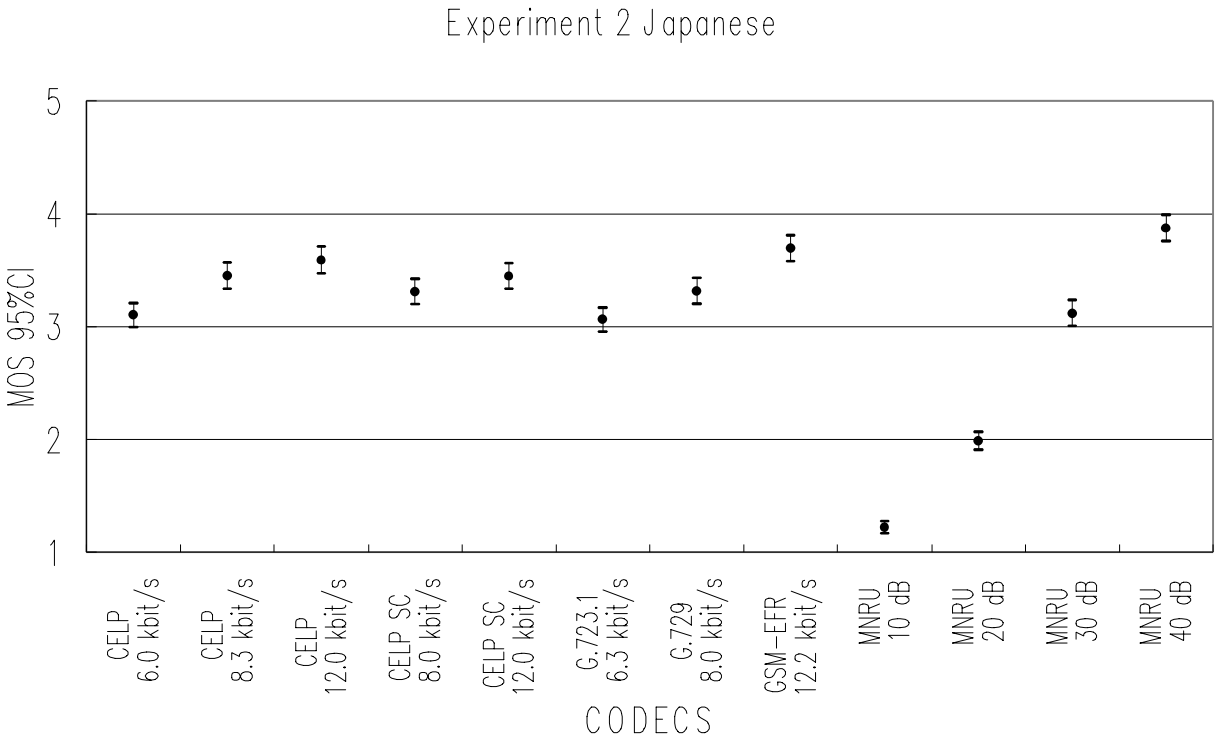


Fig. 8 MOS averaged for all Japanese items.

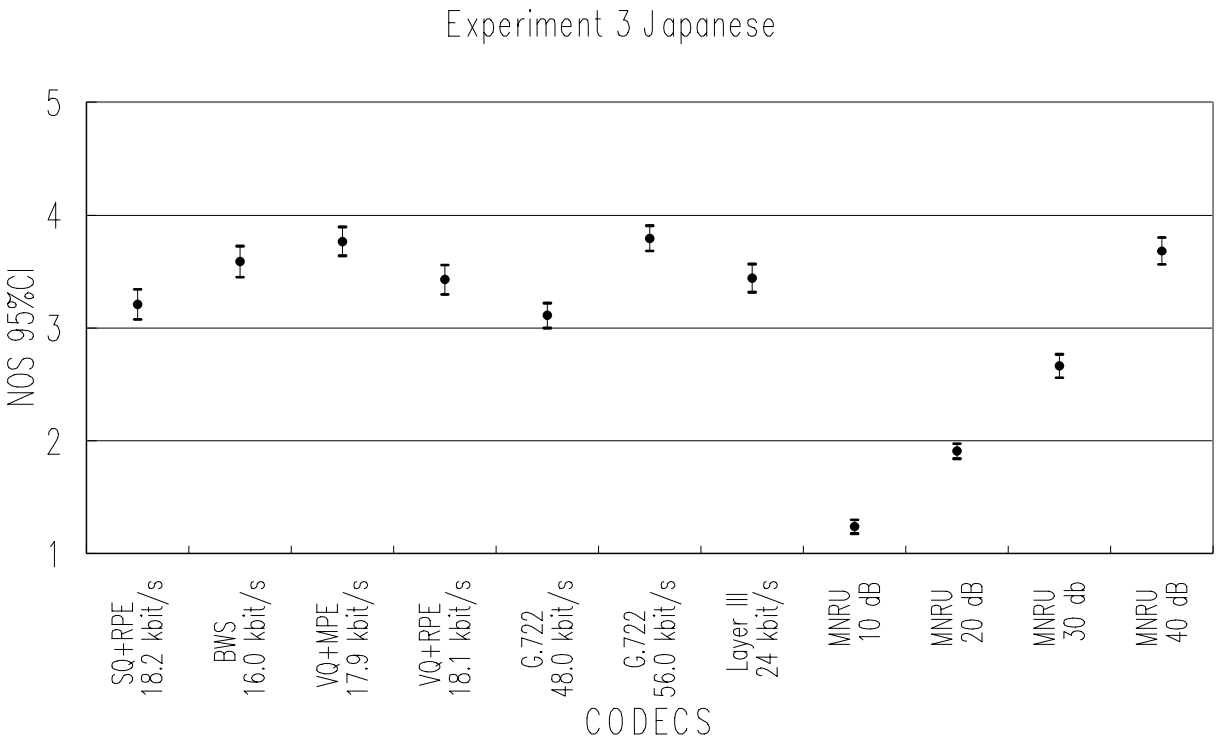


Fig. 9 MOS averaged for all Japanese items.

9.2 Performance in different languages

The results of the coder performance for each language were analysed separately to get information about the language dependency of the coders. This analysis was performed only in Nokia Research Center.

9.2.1 Nokia site

The overall performance of each coder for English, German and Swedish is shown in Tables 13, 14 and 15, and graphically in Figure 10, 11 and 12. To get reliable results, background noise items and music samples were not included into this analysis. Nevertheless, the performance in background noise and music can be verified in item by item analysis.

Performance in English language (items 07, 08, 09, 32 and 33)

Descriptives

			N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
							Lower Bound	Upper Bound
SCORE	Codec	Parametric 2 kbit/s	80	2.65	.78	.087	2.48	2.82
		Parametric 4 kbit/s	80	3.05	.98	.110	2.83	3.27
		Ref. FS1016 4.8 kbit/s	80	2.20	.75	.084	2.03	2.37
		MNRU 10	80	1.24	.51	.057	1.12	1.35
		MNRU 20	80	2.14	.76	.085	1.97	2.31
		MNRU 30	80	3.49	.80	.089	3.31	3.66
		MNRU 40	80	4.43	.63	.071	4.28	4.57
		Total	560	2.74	1.22	.051	2.64	2.84

Performance in German language (items 02, 04, 05, 26, 27, 28 and 29)

Descriptives

			N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
							Lower Bound	Upper Bound
SCORE	Codec	Parametric 2 kbit/s	112	2.29	.92	.087	2.12	2.47
		Parametric 4 kbit/s	112	2.68	1.03	.098	2.49	2.87
		Ref. FS1016 4.8 kbit/s	112	2.11	.87	.083	1.94	2.27
		MNRU 10	112	1.13	.34	.032	1.07	1.20
		MNRU 20	112	1.88	.66	.062	1.75	2.00
		MNRU 30	112	2.80	.85	.080	2.64	2.96
		MNRU 40	112	3.91	.85	.081	3.75	4.07
		Total	784	2.40	1.14	.041	2.32	2.48

Performance in Swedish language (items 136 and 138)

Descriptives

			N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
							Lower Bound	Upper Bound
SCORE	Codec	Parametric 2 kbit/s	32	3.09	.96	.170	2.75	3.44
		Parametric 4 kbit/s	32	3.53	.95	.168	3.19	3.87
		Ref. FS1016 4.8 kbit/s	32	2.50	.84	.149	2.20	2.80
		MNRU 10	32	1.19	.40	.070	1.04	1.33
		MNRU 20	32	2.03	.74	.131	1.76	2.30
		MNRU 30	32	3.53	.76	.170	3.26	3.81
		MNRU 40	32	4.66	.55	.168	4.46	4.85
		Total	224	2.93	1.30	.149	2.76	3.10

Table 13. Language dependency results of the listening test 1 (Parametric).

Performance in English language (items 06, 07, 30 and 31)

Descriptives

			N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
							Lower Bound	Upper Bound
SCORE	Codec	CELP (Mode VIII multi rate) 6 kbit/s	64	2.78	.79	.098	2.58	2.98
		CELP (Mode VIII multi rate) 8.3 kbit/s	64	3.31	.97	.122	3.07	3.56
		CELP (Mode VIII multi rate) 12 kbit/s	64	3.55	1.01	.126	3.30	3.80
		CELP (Mode VIII scaleable) 8 kbit/s	64	2.97	.80	.100	2.77	3.17
		CELP (Mode VIII scaleable) 12 kbit/s	64	3.25	.91	.114	3.02	3.48
		Ref. ITU-T G.723.1 6.3 kbit/s	64	2.92	.90	.112	2.70	3.15
		Ref. ITU-T G.729 8 kbit/s	64	3.50	.87	.109	3.28	3.72
		Ref. GSM-EFR 12.2 kbit/s	64	4.03	.91	.113	3.80	4.26
		MNRU 10	64	1.09	.29	.037	1.02	1.17
		MNRU 20	64	1.78	.70	.088	1.61	1.96
		MNRU 30	64	3.11	.99	.124	2.86	3.36
		MNRU 40	64	3.92	.93	.116	3.69	4.15
		Total	768	3.02	1.17	.042	2.94	3.10

Performance in German language (items 02, 04, 05, 26, 27 and 29)

Descriptives

			N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
							Lower Bound	Upper Bound
SCORE	Codec	CELP (Mode VIII multi rate) 6 kbit/s	96	2.58	.79	.081	2.42	2.74
		CELP (Mode VIII multi rate) 8.3 kbit/s	96	3.10	.79	.080	2.94	3.26
		CELP (Mode VIII multi rate) 12 kbit/s	96	3.69	.85	.087	3.52	3.86
		CELP (Mode VIII scaleable) 8 kbit/s	96	2.92	.71	.072	2.77	3.06
		CELP (Mode VIII scaleable) 12 kbit/s	96	3.29	.79	.081	3.13	3.45
		Ref. ITU-T G.723.1 6.3 kbit/s	96	2.96	.74	.075	2.81	3.11
		Ref. ITU-T G.729 8 kbit/s	96	3.21	.75	.077	3.06	3.36
		Ref. GSM-EFR 12.2 kbit/s	96	3.70	.77	.079	3.54	3.85
		MNRU 10	96	1.13	.33	.034	1.06	1.19
		MNRU 20	96	1.61	.62	.063	1.49	1.74
		MNRU 30	96	2.55	.77	.078	2.40	2.71
		MNRU 40	96	3.70	.82	.084	3.53	3.86
		Total	1152	2.87	1.07	.031	2.81	2.93

Performance in Swedish language (items 136 and 138)

Descriptives

			N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
							Lower Bound	Upper Bound
SCORE	Codec	CELP (Mode VIII multi rate) 6 kbit/s	32	2.97	.74	.131	2.70	3.24
		CELP (Mode VIII multi rate) 8.3 kbit/s	32	3.53	.72	.127	3.27	3.79
		CELP (Mode VIII multi rate) 12 kbit/s	32	4.16	.72	.128	3.90	4.42
		CELP (Mode VIII scaleable) 8 kbit/s	32	3.09	.69	.122	2.85	3.34
		CELP (Mode VIII scaleable) 12 kbit/s	32	3.50	.84	.149	3.20	3.80
		Ref. ITU-T G.723.1 6.3 kbit/s	32	3.22	.61	.108	3.00	3.44
		Ref. ITU-T G.729 8 kbit/s	32	3.50	.72	.127	3.24	3.76
		Ref. GSM-EFR 12.2 kbit/s	32	4.31	.59	.105	4.10	4.53
		MNRU 10	32	1.16	.37	.065	1.02	1.29
		MNRU 20	32	2.03	.74	.131	1.76	2.30
		MNRU 30	32	3.50	.92	.162	3.17	3.83
		MNRU 40	32	4.50	.51	.090	4.32	4.68
		Total	384	3.29	1.13	.058	3.18	3.40

Table 14. Language dependency results of the listening test 2 (NB-CELP).

Performance in English language (items 06, 07, 30 and 33)

Descriptives

			N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
							Lower Bound	Upper Bound
SCORE	Codec	CELP (fr III) 18 .2	64	2.83	.95	.119	2.59	3.07
		CELP (BW sc) 16	64	3.25	.91	.114	3.02	3.48
		Opt VQ+MPE 17.9	64	3.64	1.03	.129	3.38	3.90
		Opt VQ+RPE 18.1	64	3.38	1.02	.127	3.12	3.63
		Ref G722 48	64	3.17	.88	.110	2.95	3.39
		Ref G722 56	64	3.69	1.08	.135	3.42	3.96
		Ref MPEG-2 III 24	64	3.14	.92	.115	2.91	3.37
		MNRU 10	64	1.11	.36	.045	1.02	1.20
		MNRU 20	64	1.67	.71	.089	1.49	1.85
		MNRU 30	64	2.50	.82	.102	2.30	2.70
		MNRU 40	64	3.47	1.13	.141	3.19	3.75
		Total	704	2.89	1.20	.045	2.81	2.98

Performance in German language (items 02, 04, 28 and 29)

Descriptives

			N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
							Lower Bound	Upper Bound
SCORE	Codec	CELP (fr III) 18 .2	64	3.47	.89	.111	3.25	3.69
		CELP (BW sc) 16	64	3.41	.77	.096	3.21	3.60
		Opt VQ+MPE 17.9	64	3.83	.83	.103	3.62	4.03
		Opt VQ+RPE 18.1	64	3.97	.73	.092	3.79	4.15
		Ref G722 48	64	3.50	1.02	.128	3.24	3.76
		Ref G722 56	64	4.05	.74	.093	3.86	4.23
		Ref MPEG-2 III 24	64	3.34	.93	.116	3.11	3.58
		MNRU 10	64	1.11	.31	.039	1.03	1.19
		MNRU 20	64	1.63	.65	.082	1.46	1.79
		MNRU 30	64	2.56	.66	.083	2.40	2.73
		MNRU 40	64	3.77	.94	.117	3.53	4.00
		Total	704	3.15	1.22	.046	3.06	3.24

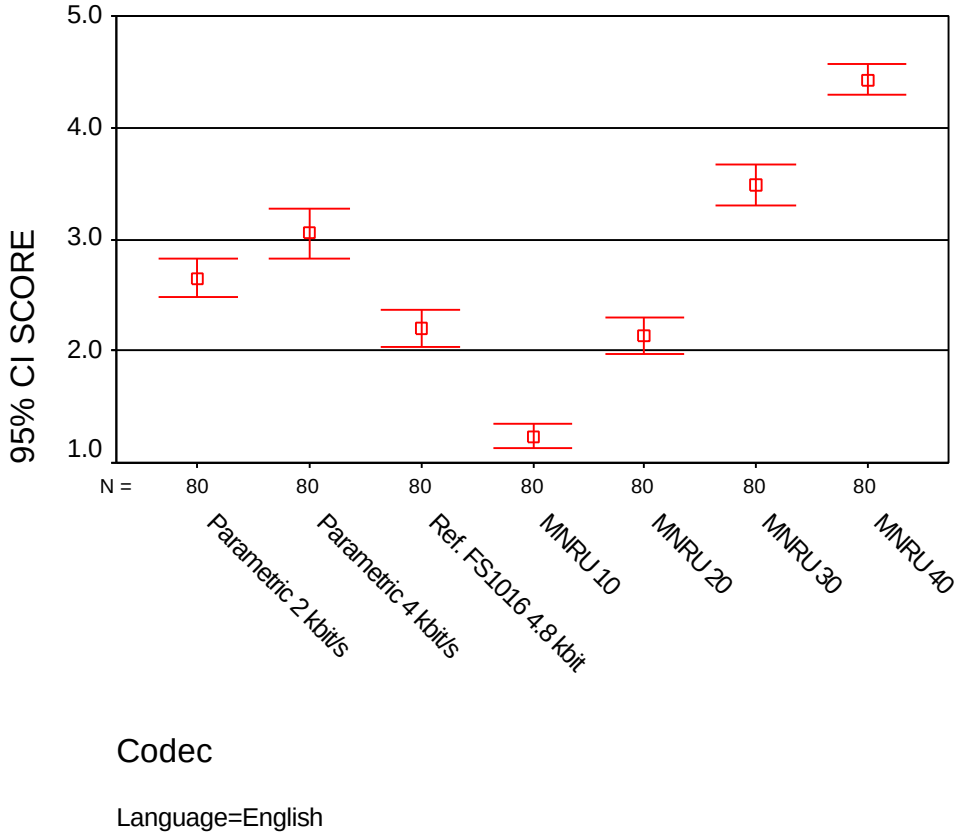
Performance in Swedish language (items 136 and 138)

Descriptives

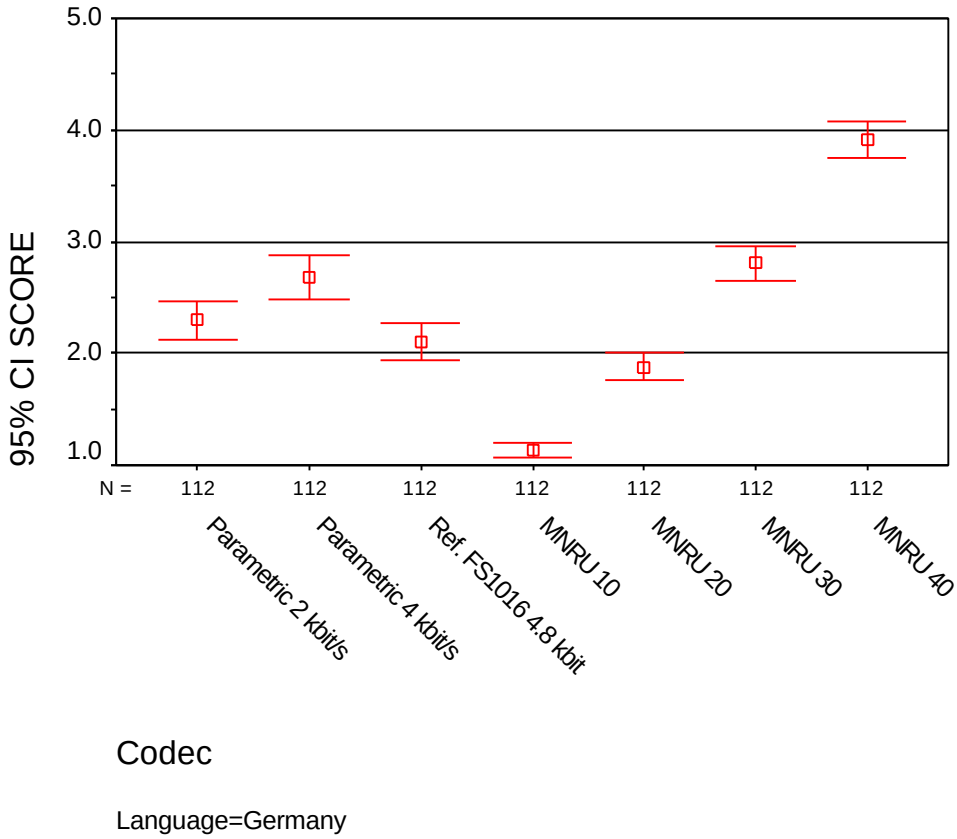
			N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean	
							Lower Bound	Upper Bound
SCORE	Codec	CELP (fr III) 18 .2	32	3.50	.84	.149	3.20	3.80
		CELP (BW sc) 16	32	3.94	.72	.127	3.68	4.20
		Opt VQ+MPE 17.9	32	4.00	.88	.156	3.68	4.32
		Opt VQ+RPE 18.1	32	3.88	.75	.133	3.60	4.15
		Ref G722 48	32	3.97	.78	.138	3.69	4.25
		Ref G722 56	32	4.41	.67	.118	4.17	4.65
		Ref MPEG-2 III 24	32	3.34	.79	.139	3.06	3.63
		MNRU 10	32	1.13	.34	.059	1.00	1.25
		MNRU 20	32	1.84	.68	.120	1.60	2.09
		MNRU 30	32	2.97	.78	.138	2.69	3.25
		MNRU 40	32	4.34	.65	.115	4.11	4.58
		Total	352	3.39	1.23	.065	3.26	3.52

Table 15. Language dependency results of the listening test 3 (WB-CELP).

Performance in English language (items 07, 08, 09, 32 and 33)



Performance in German language (items 02, 04, 05, 26, 27, 28 and 29)



Performance in Swedish language (items 136 and 138)

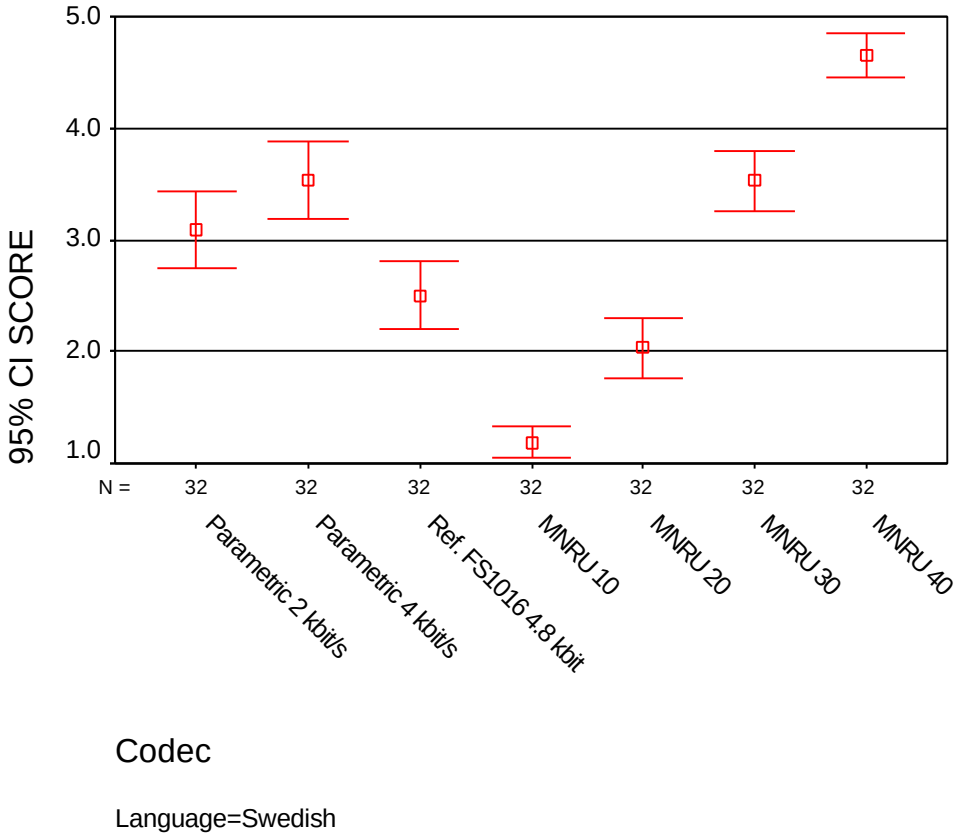
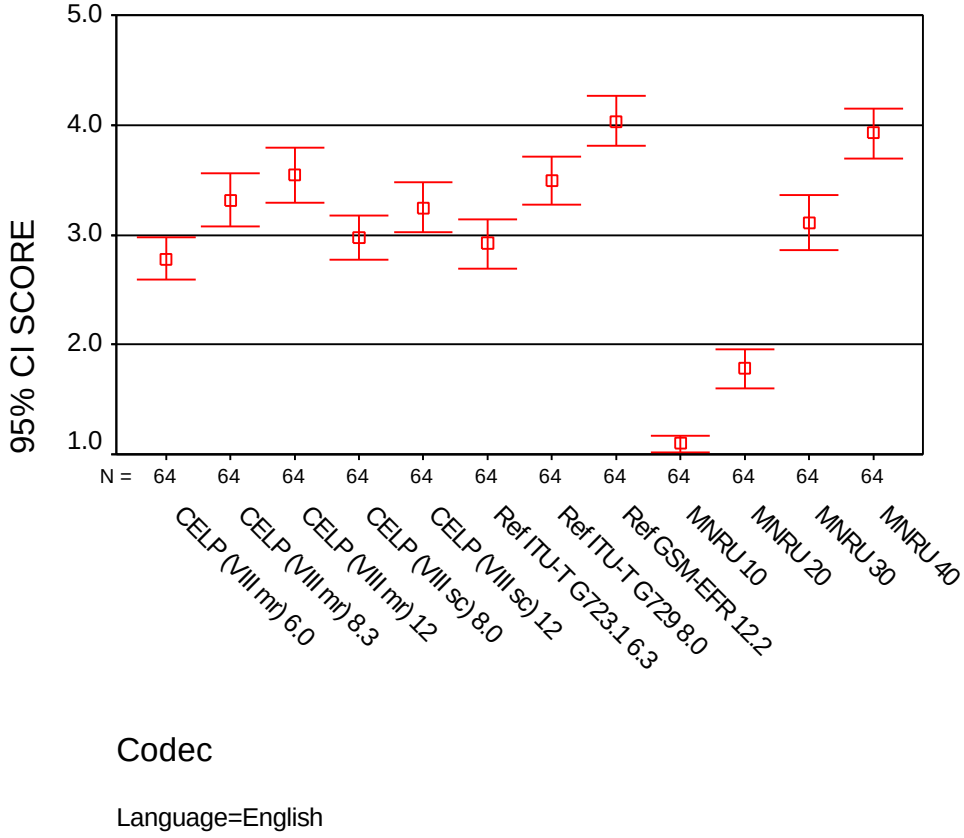
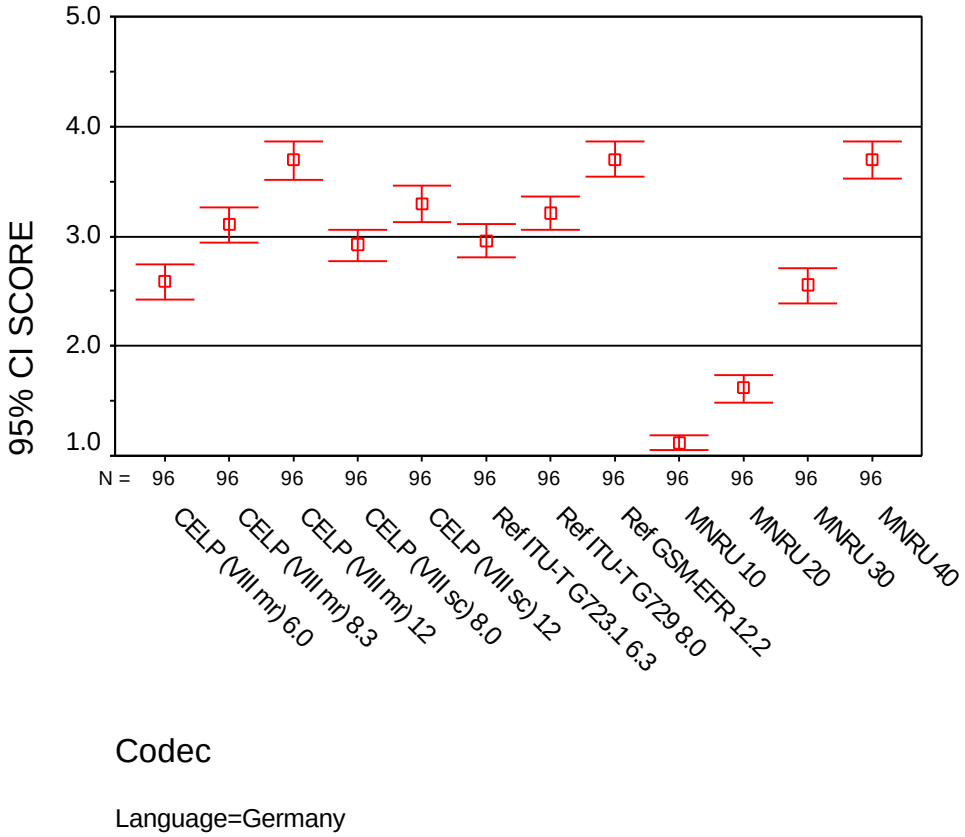


Figure 10. Results of the listening test 1 (Parametric).

Performance in English language (items 06, 07, 30 and 31)



Performance in German language (items 02, 04, 05, 26, 27 and 29)



Performance in Swedish language (items 136 and 138)

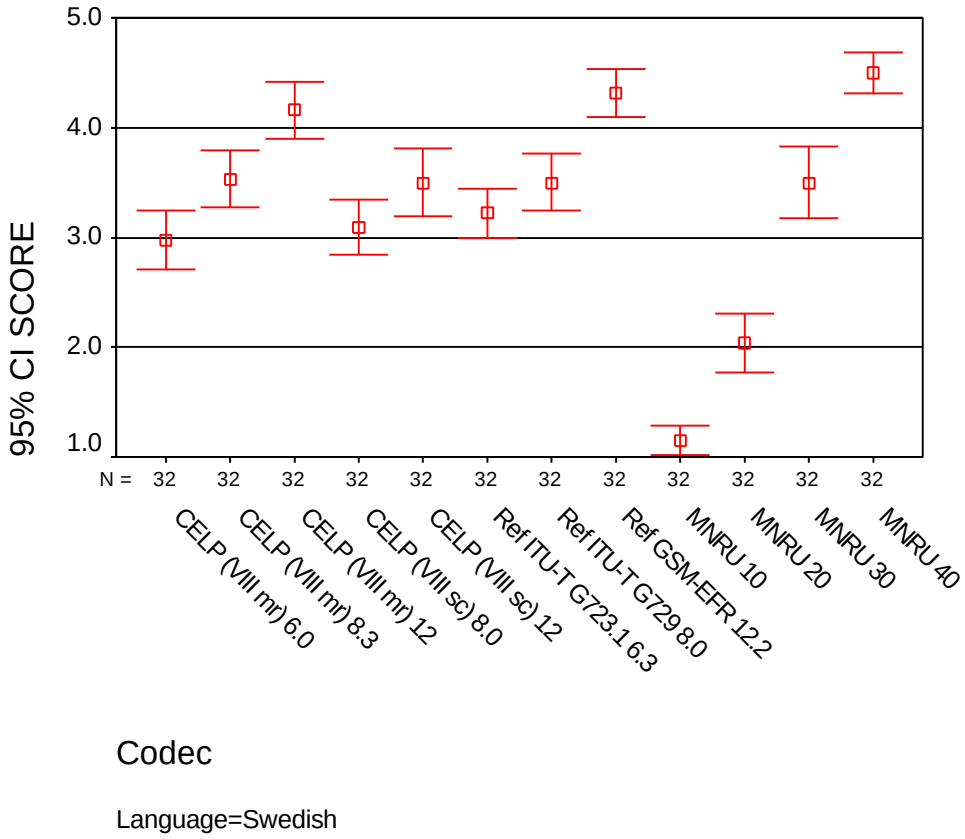
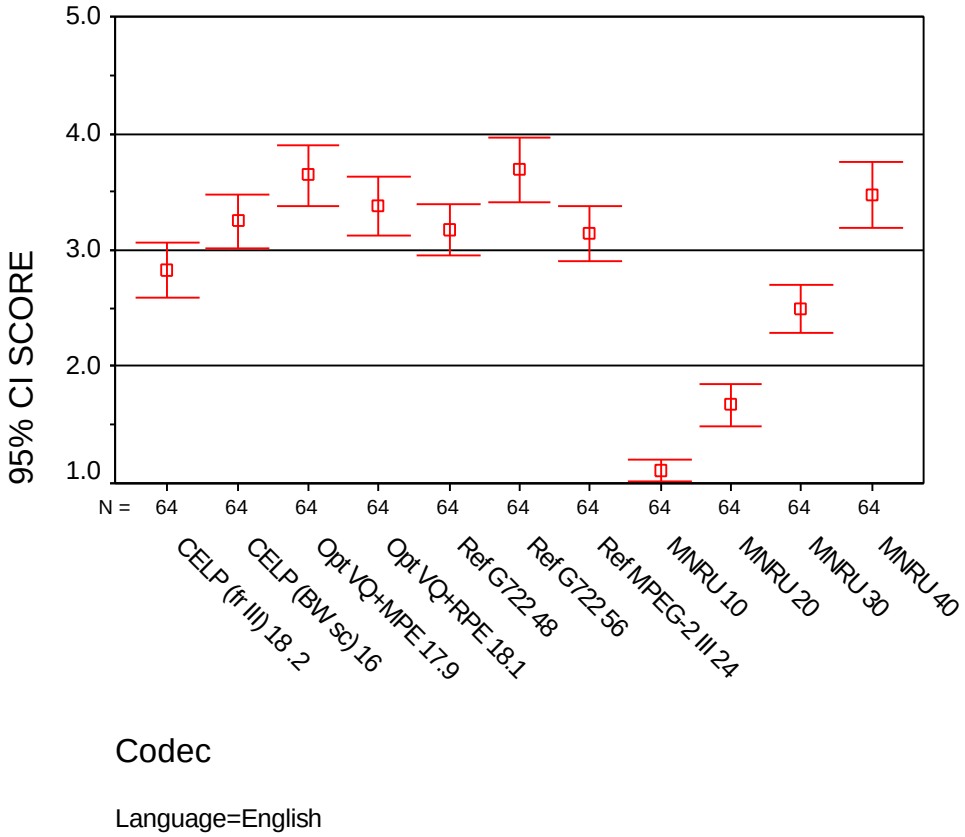
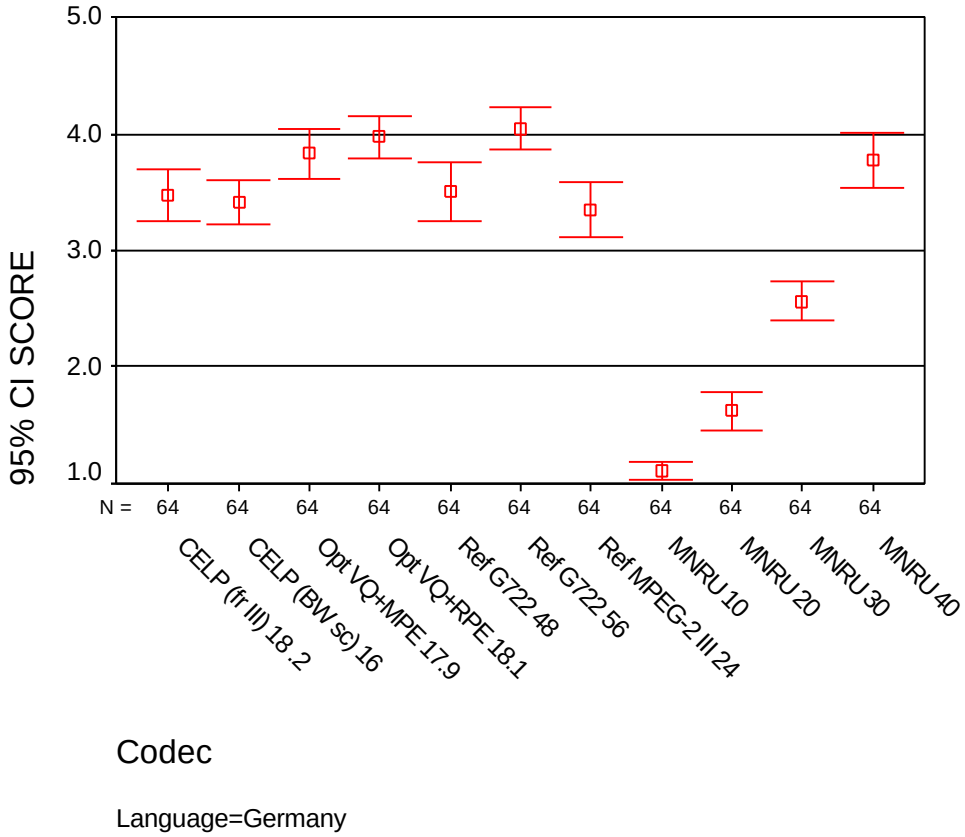


Figure 11. Results of the listening test 2 (NB-CELP).

Performance in English language (items 06, 07, 30 and 33)



Performance in German language (items 02, 04, 28 and 29)



Performance in Swedish language (items 136 and 138)

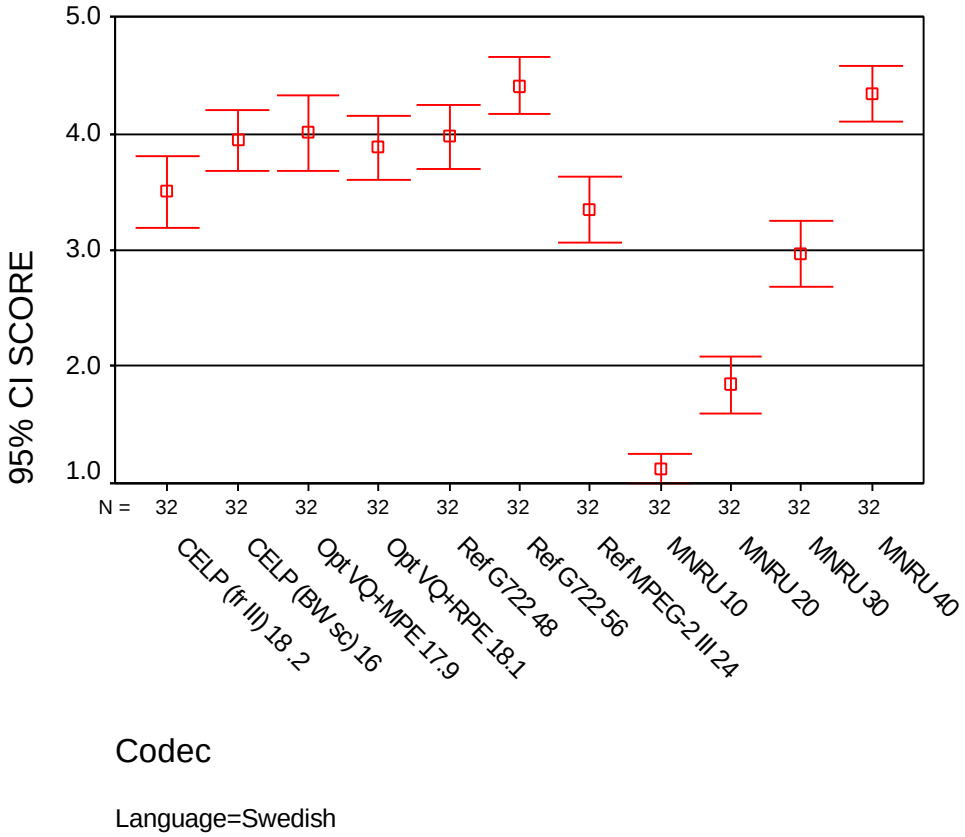


Figure 12. Results of the listening test 3 (WB-CELP).

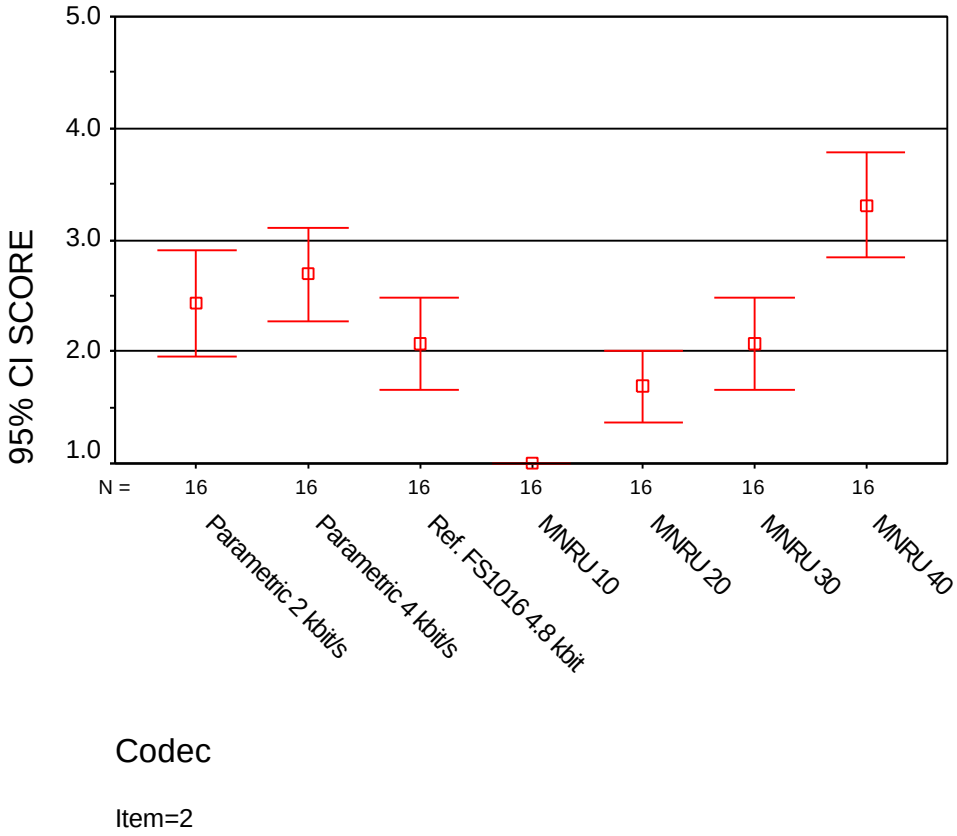
9.3 Performance item by item

The results of the coder performance for each test item were analysed separately.

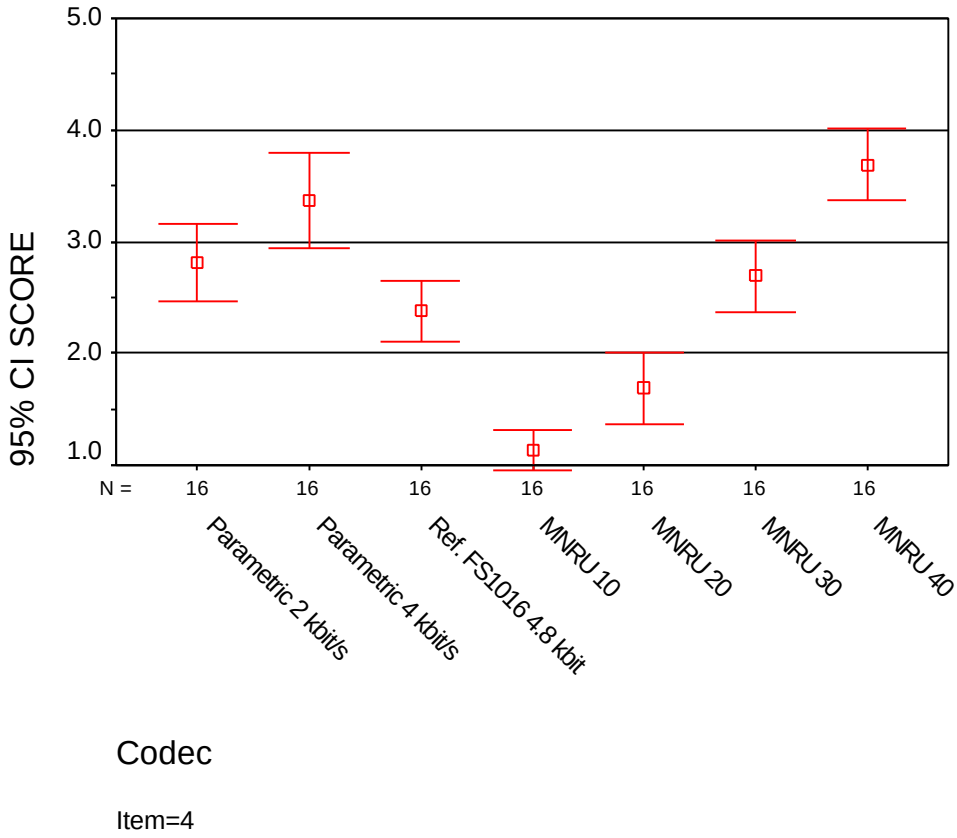
9.3.1 Nokia site

The performance of each coder in experiments 1, 2 and 3 are shown graphically in Figures 13, 14 and 15, respectively.

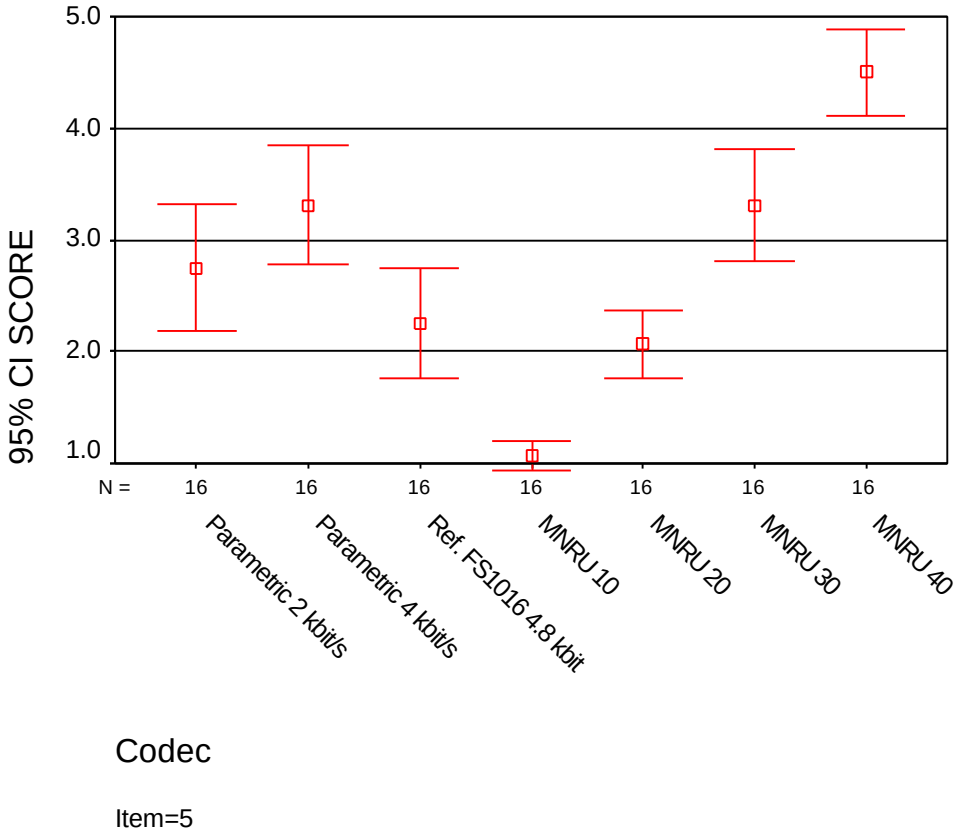
Item 02, Male (German)



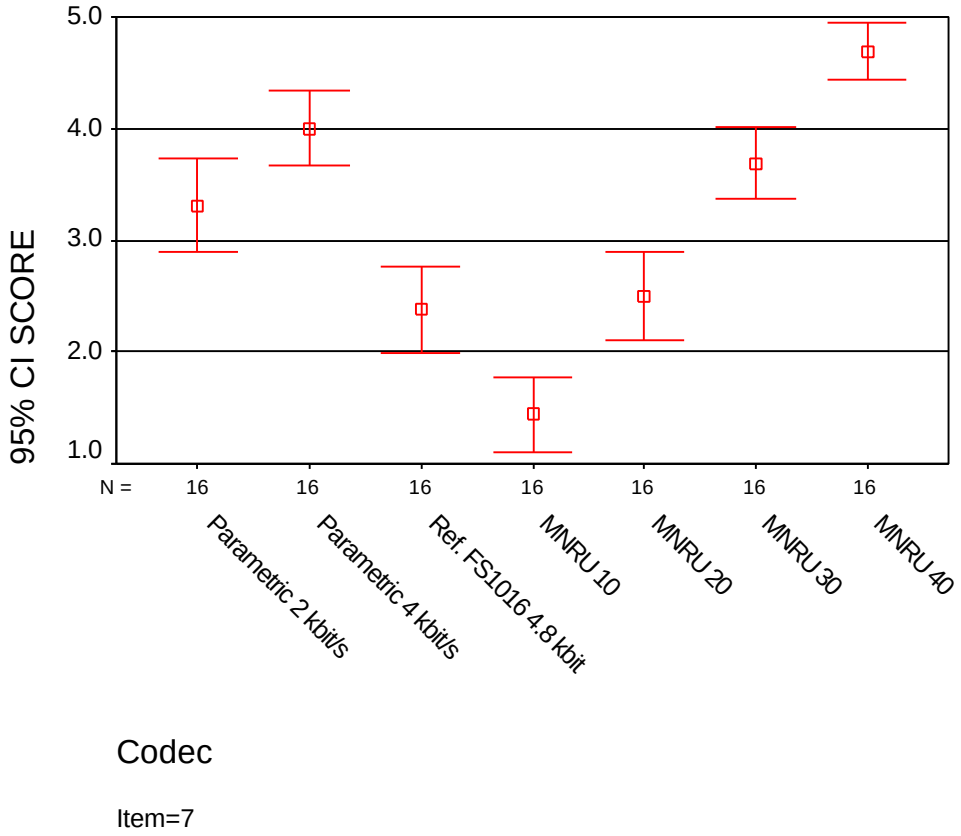
Item 04, Male (German)



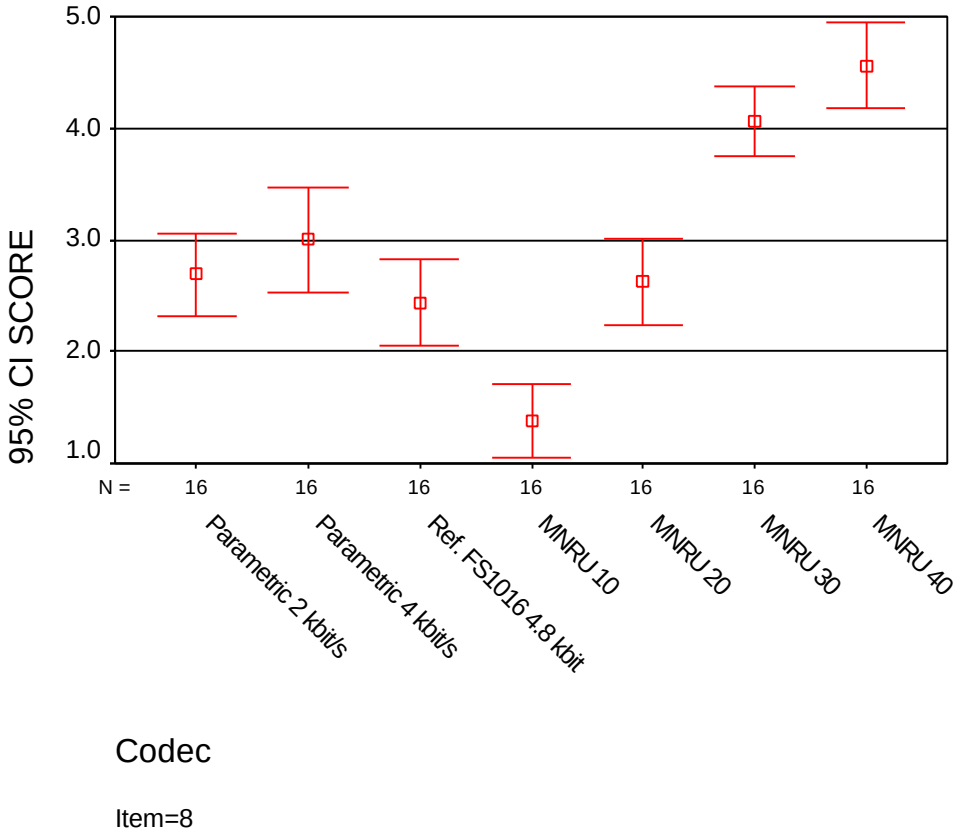
Item 05, Male (German)



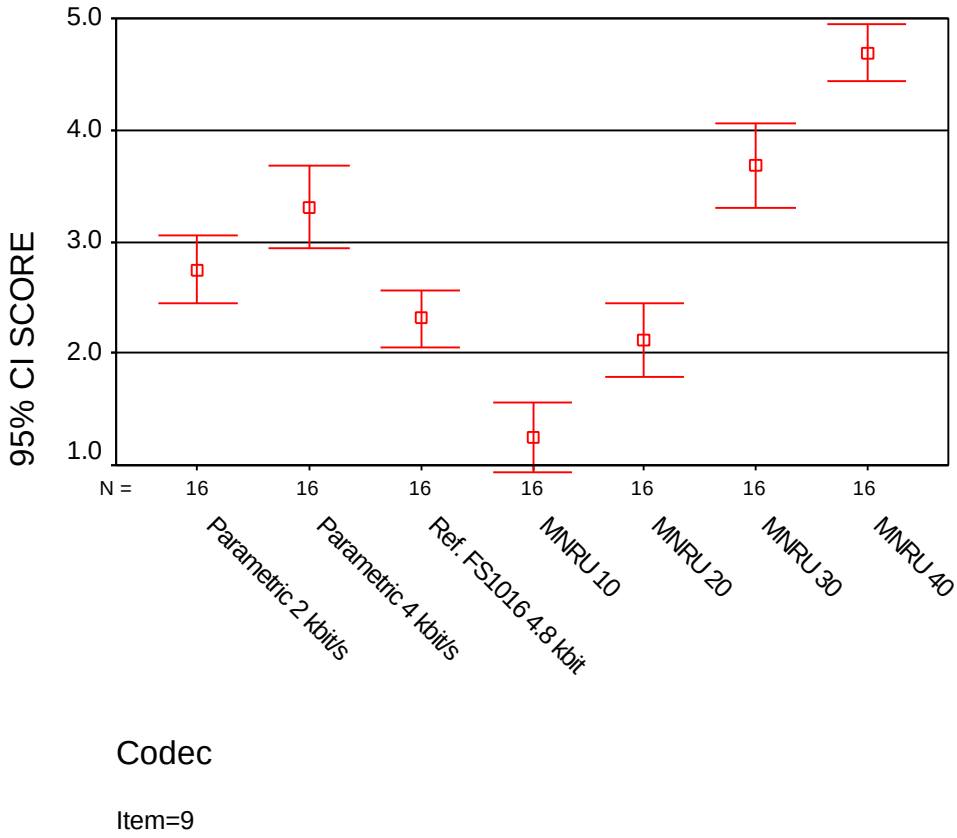
Item 07, Male (English)



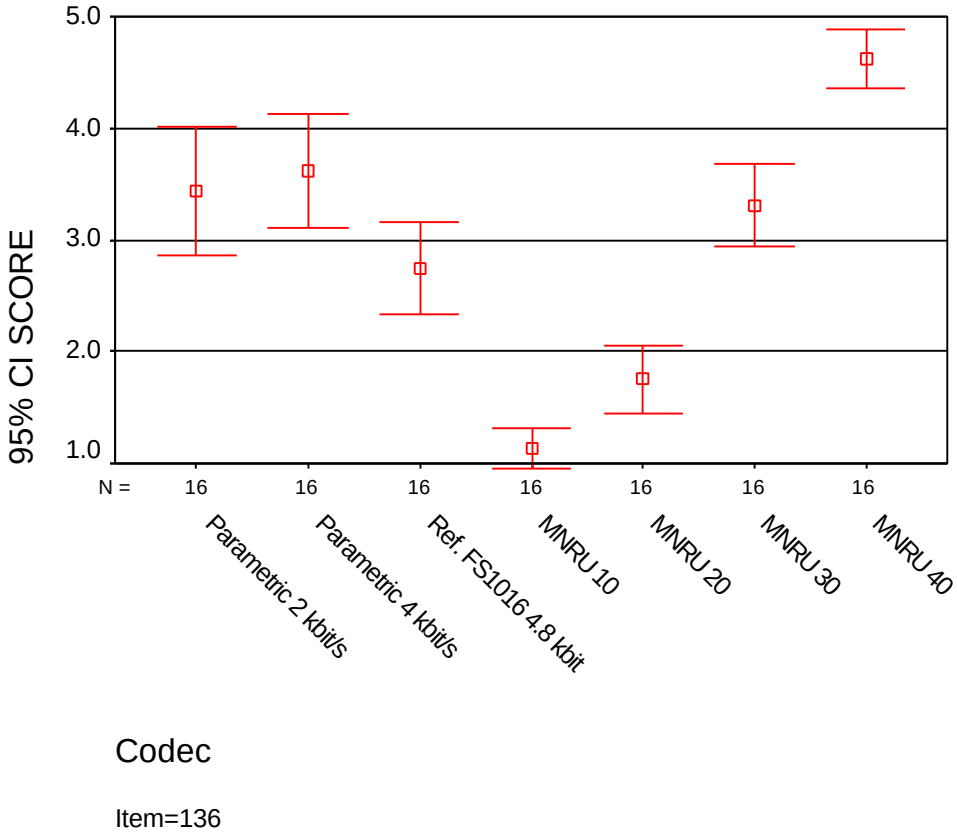
Item 08, Male (English)



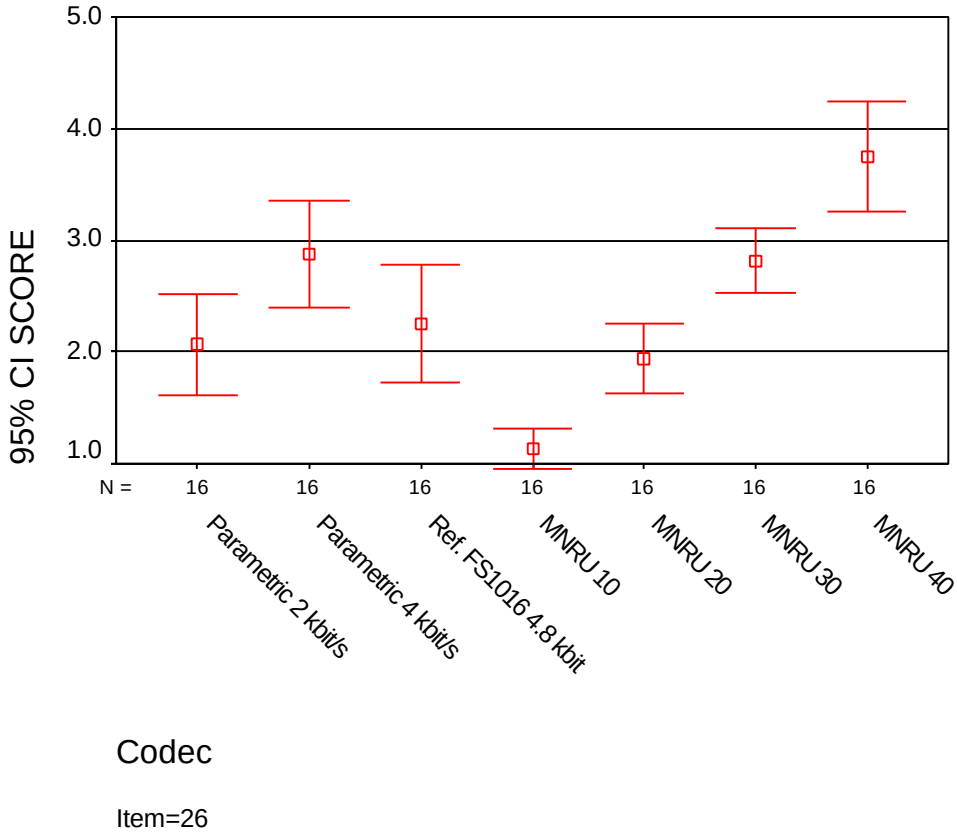
Item 09, Male (English)



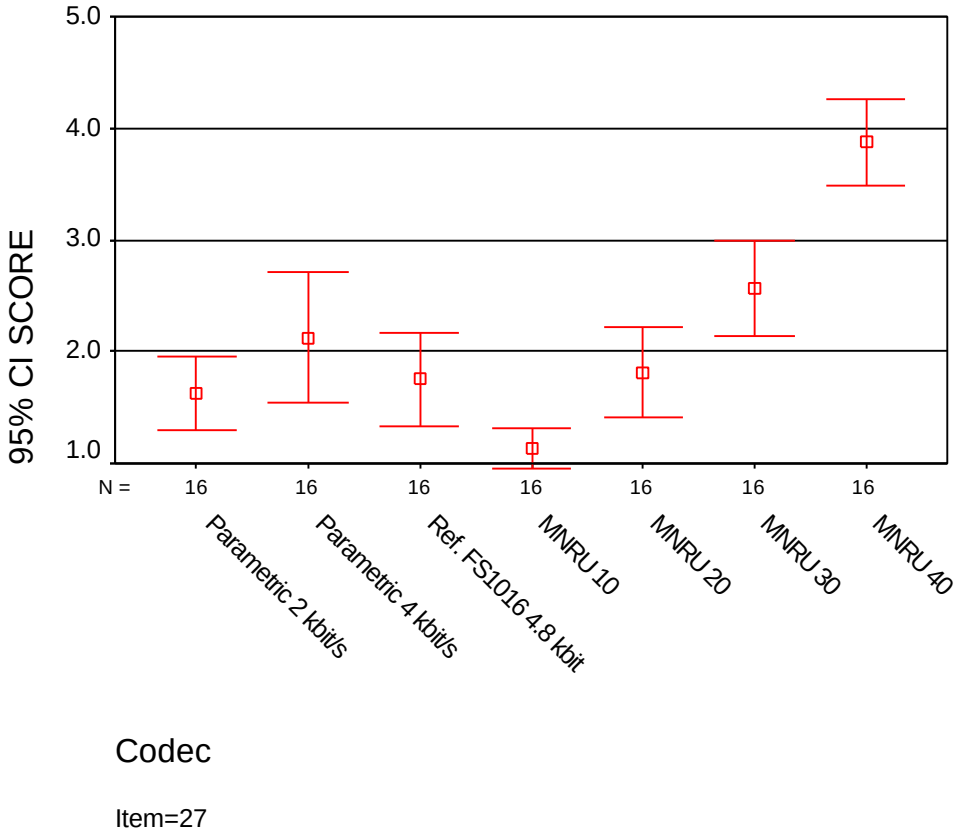
Item 136, Male (Swedish)



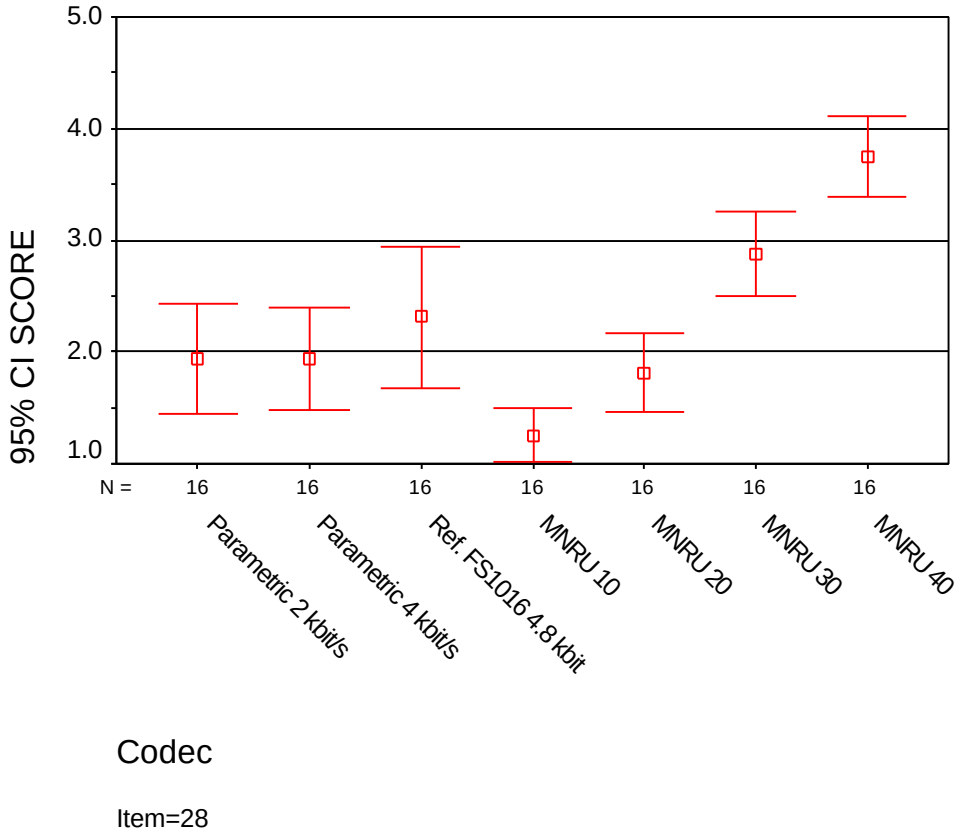
Item 26, Female (German)



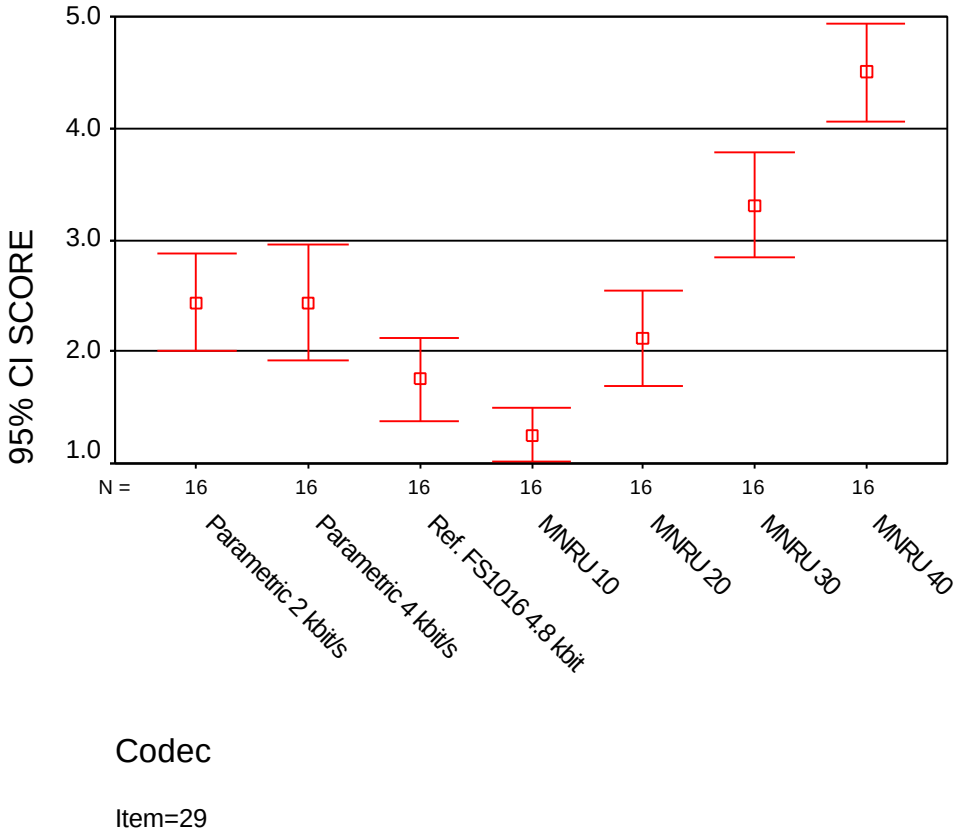
Item 27, Female (German)



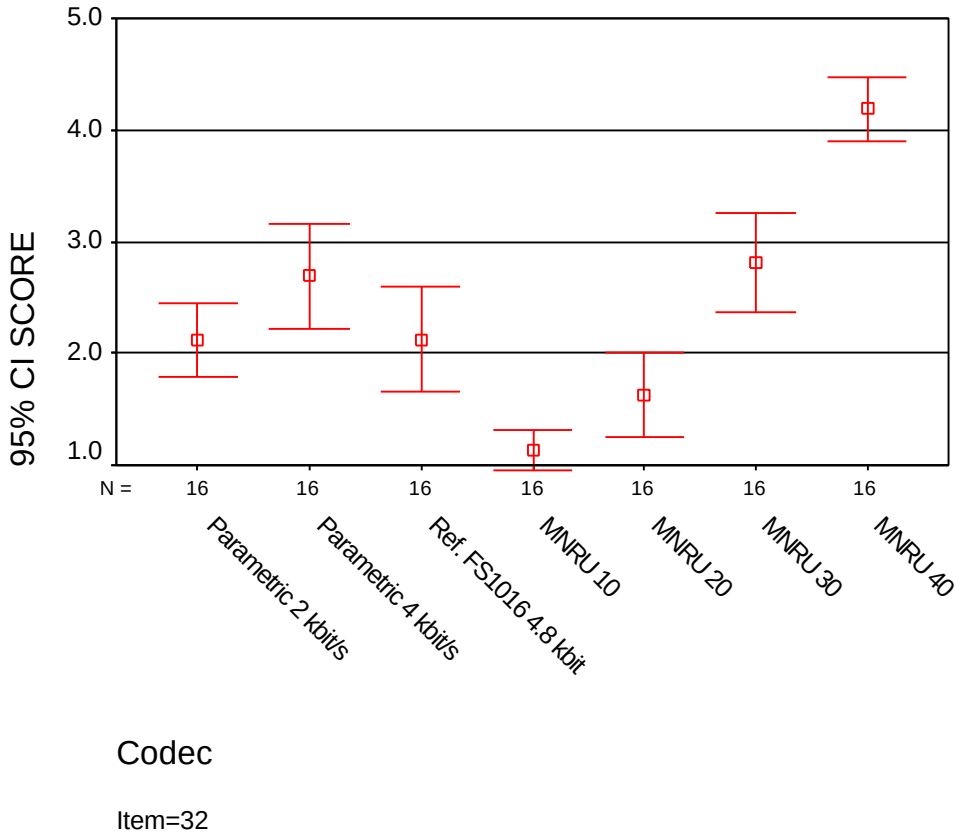
Item 28, Female (German)



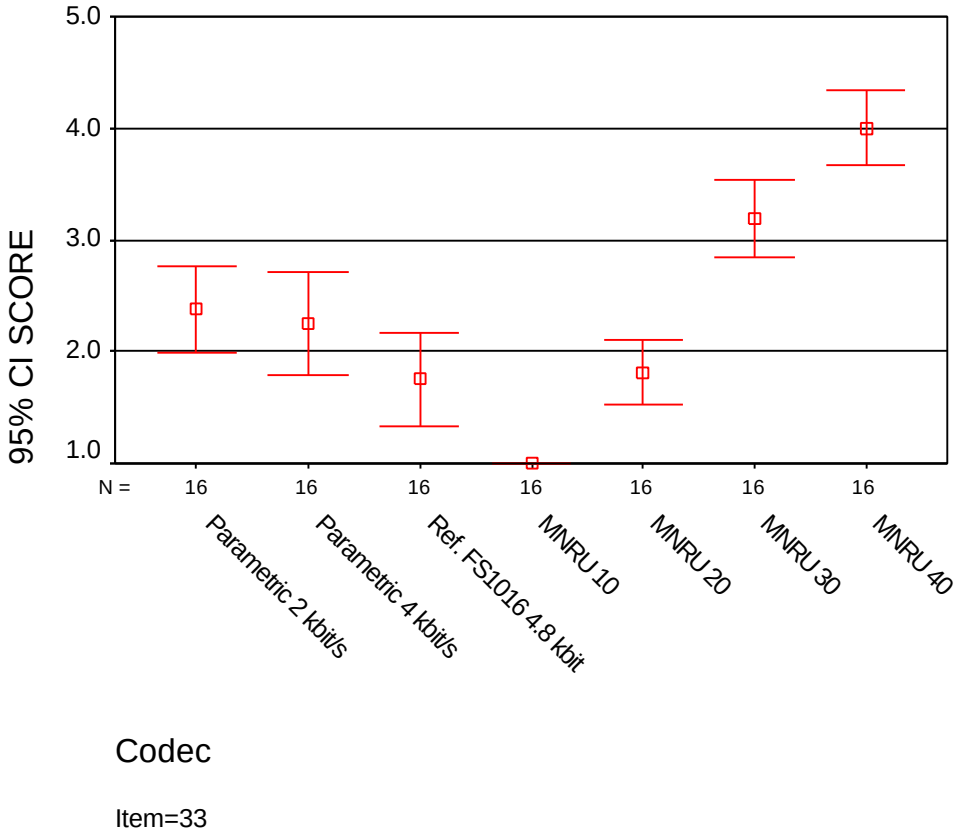
Item 29, Female (German)



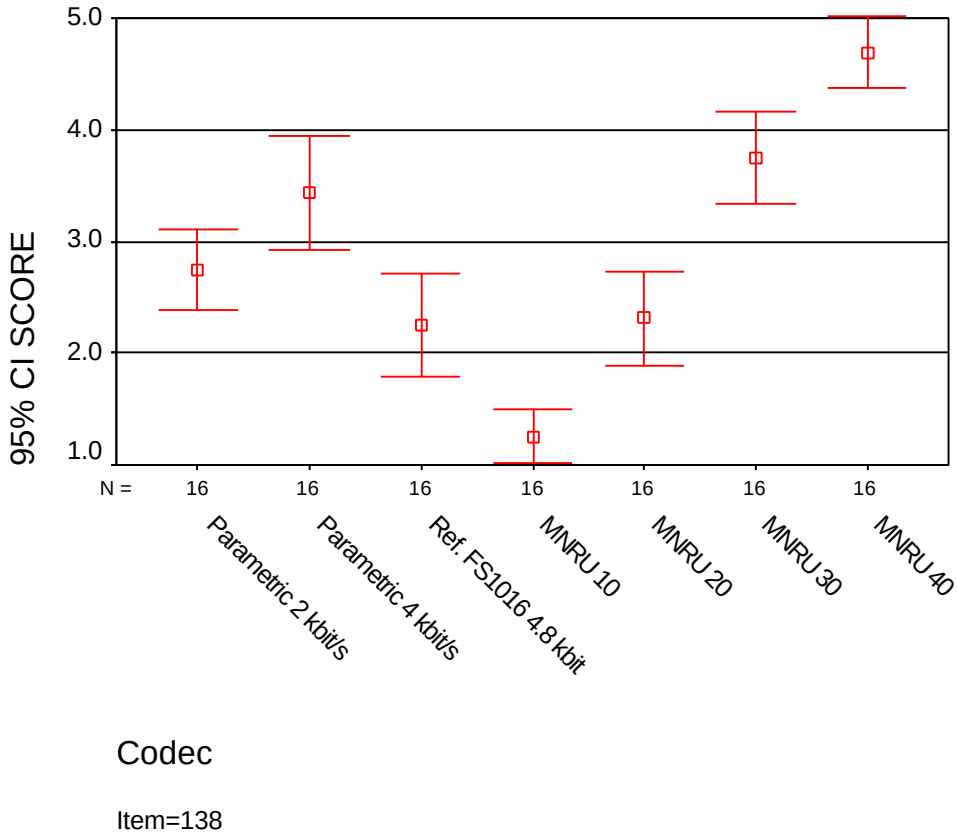
Item 32, Female (English)



Item 33, Female (English)



Item 138, Female (Swedish)



Item 08_c1, Male Car background noise (English)

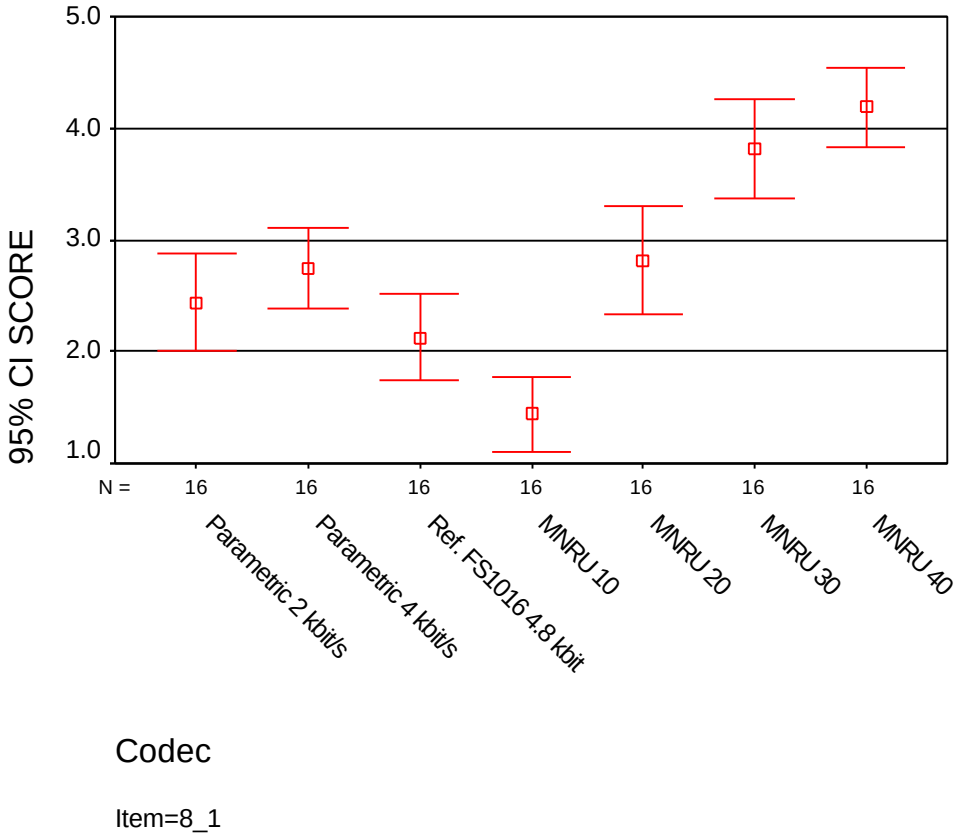
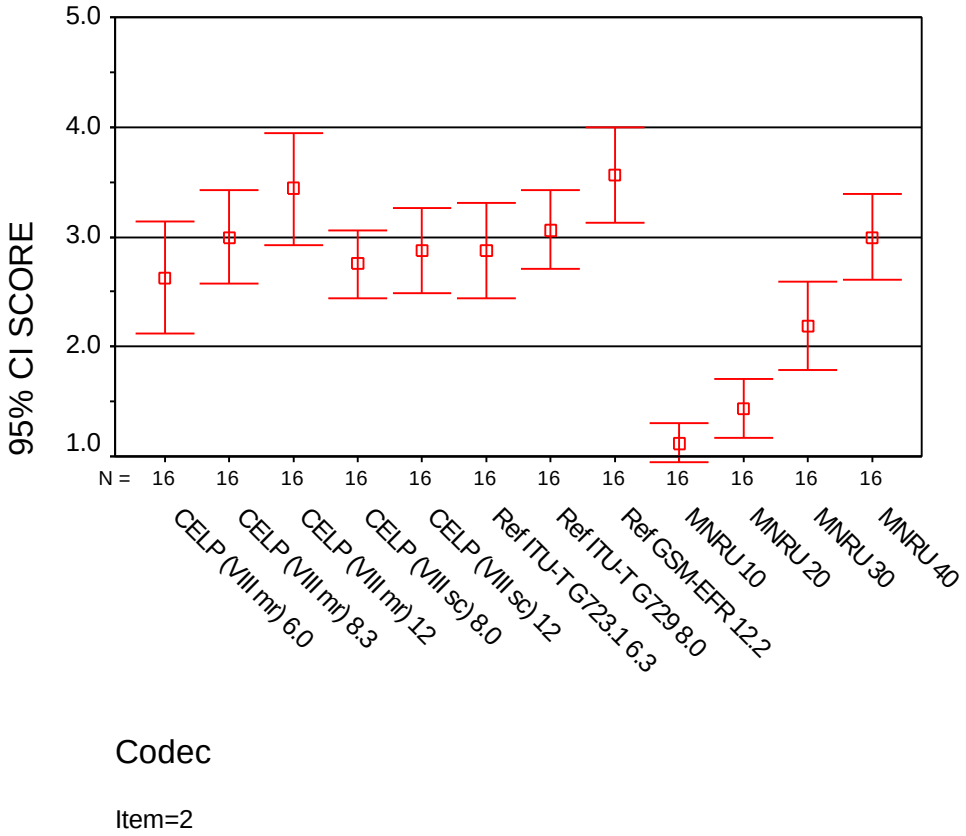
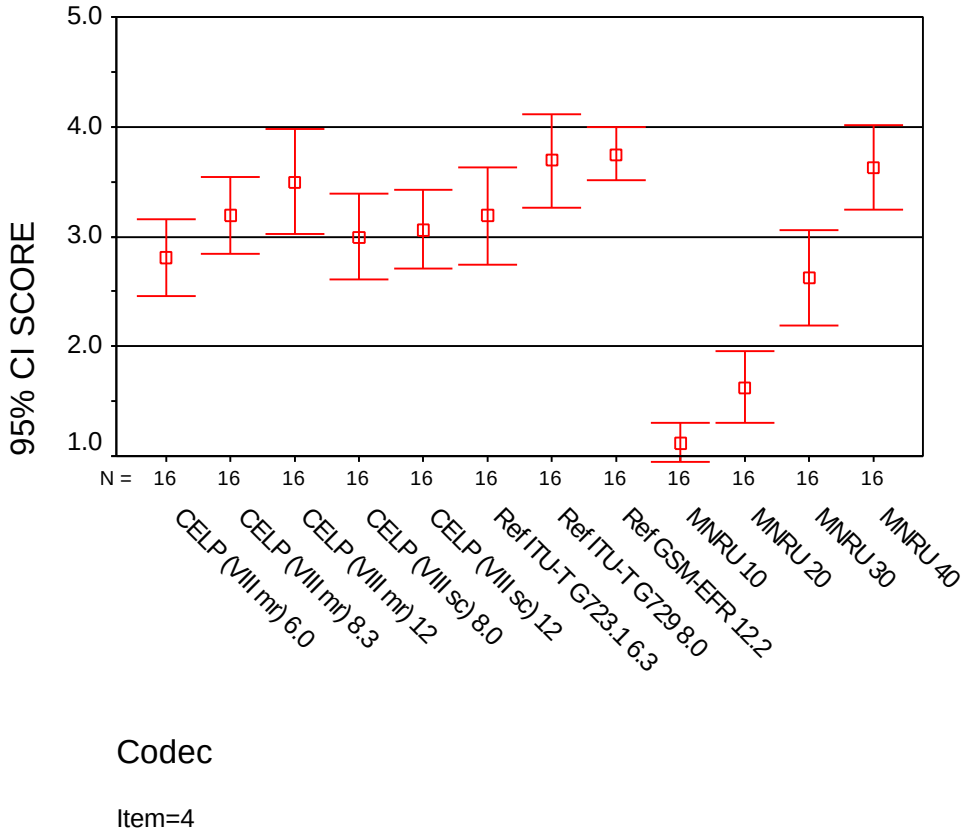


Figure 13. Results of the listening test 1 (Parametric).

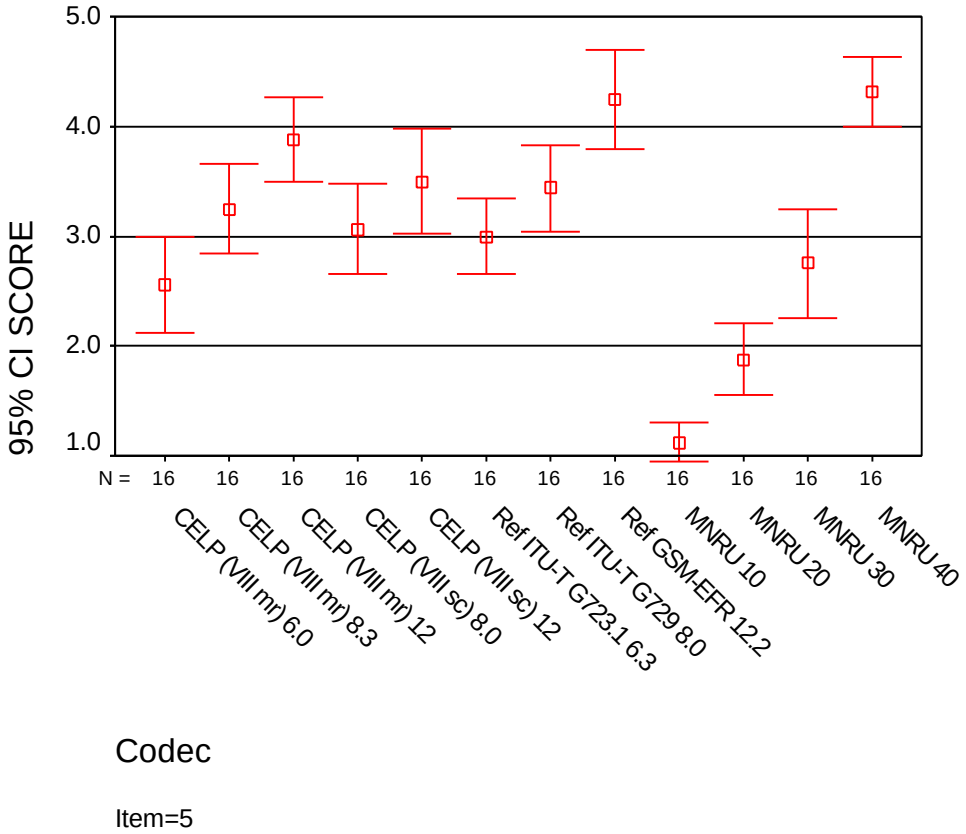
Item 02, Male (German)



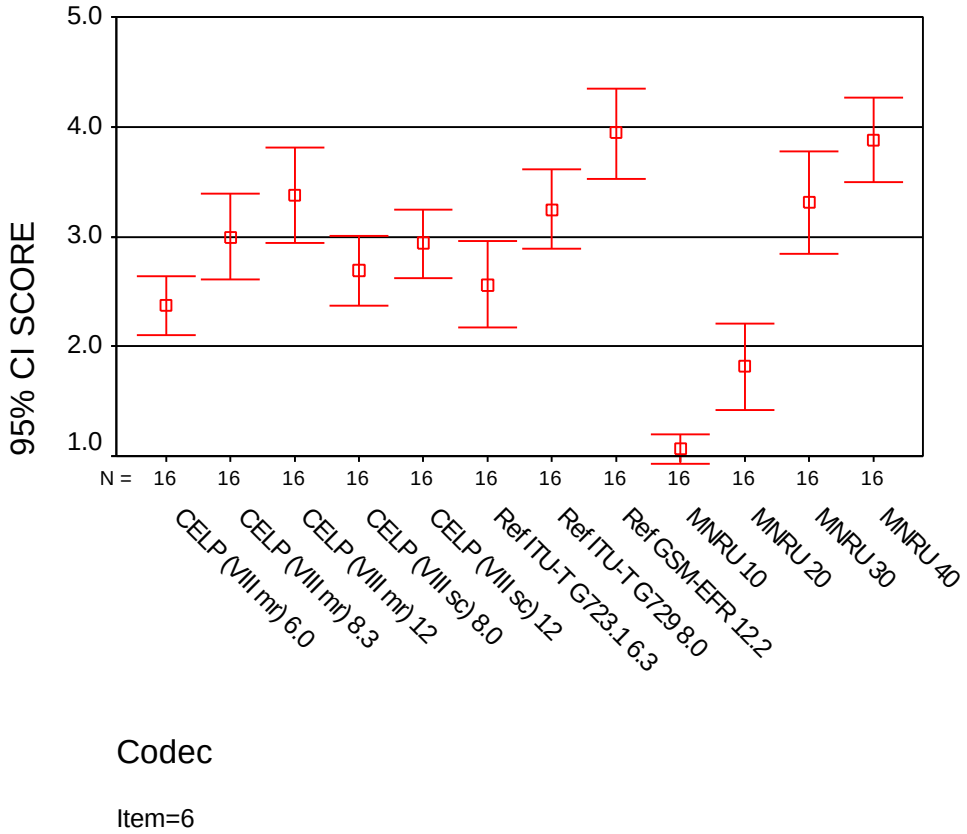
Item 04, Male (German)



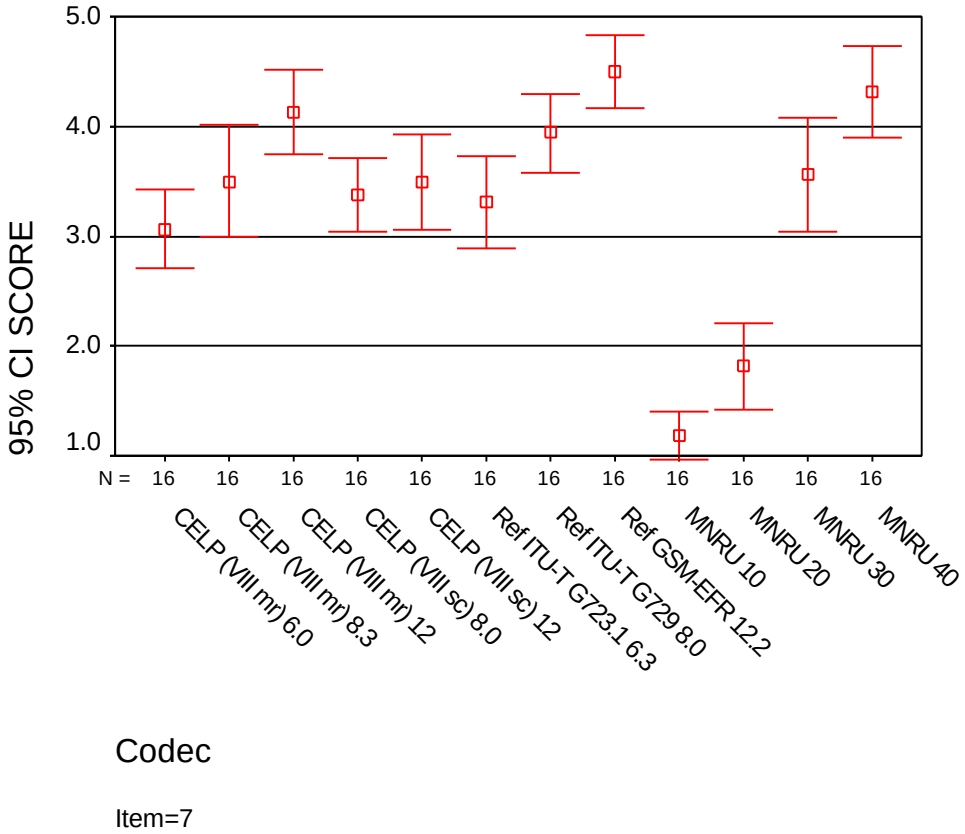
Item 05, Male (German)



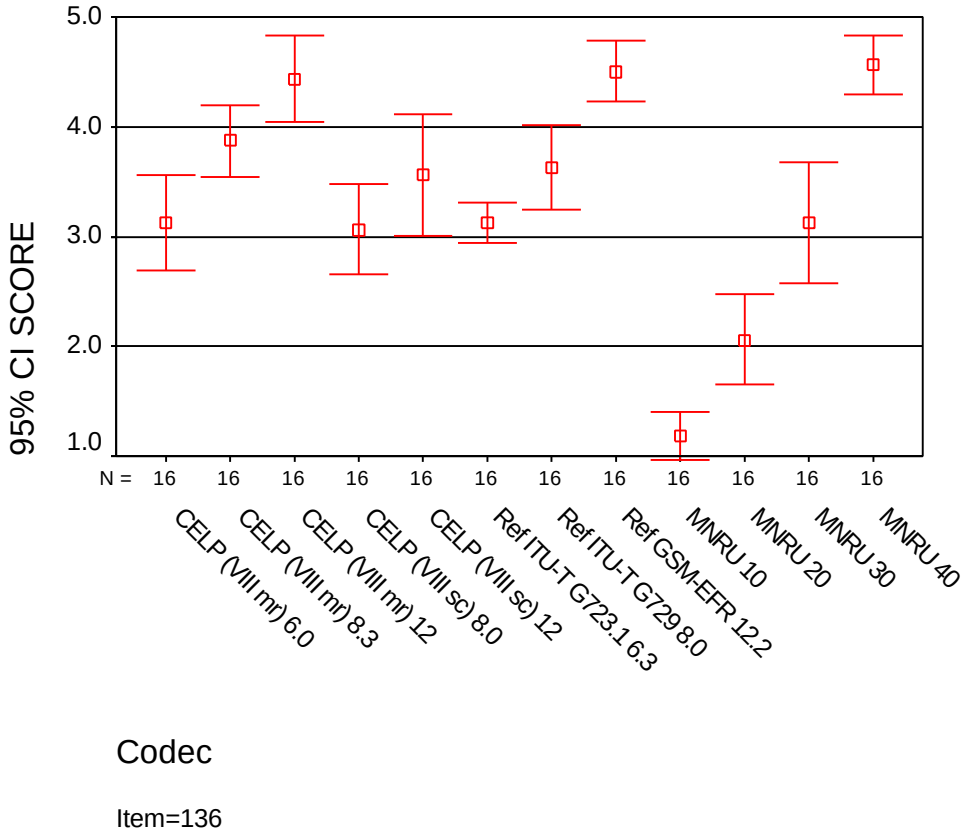
Item 06, Male (English)



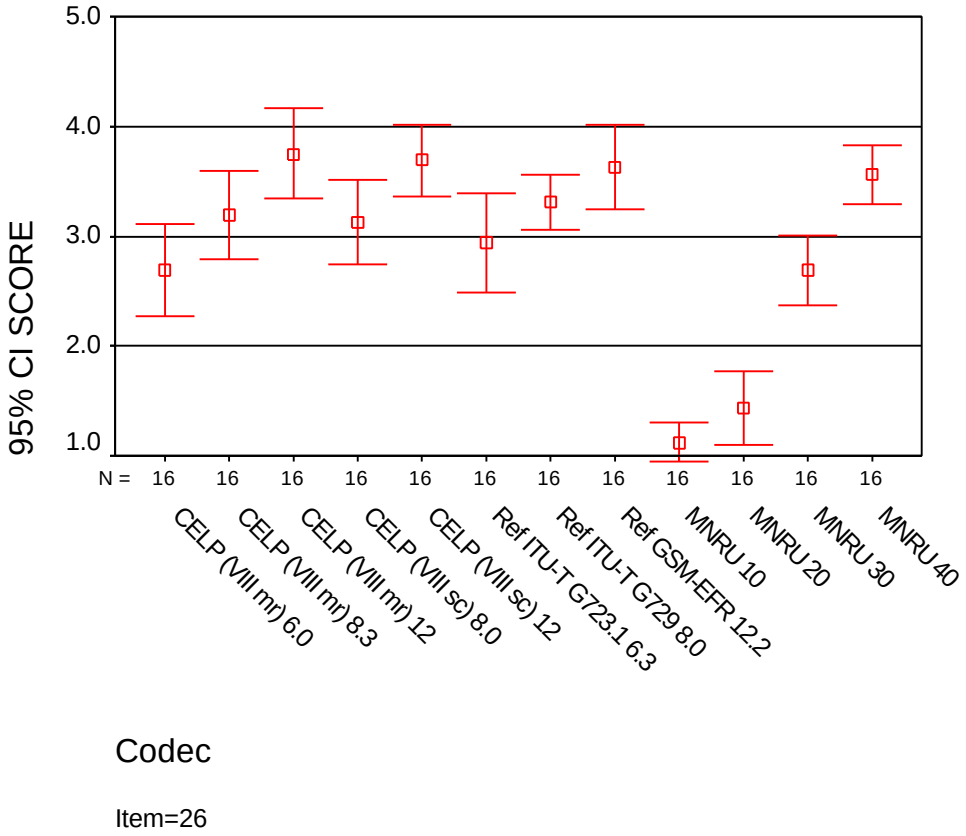
Item 07, Male (English)



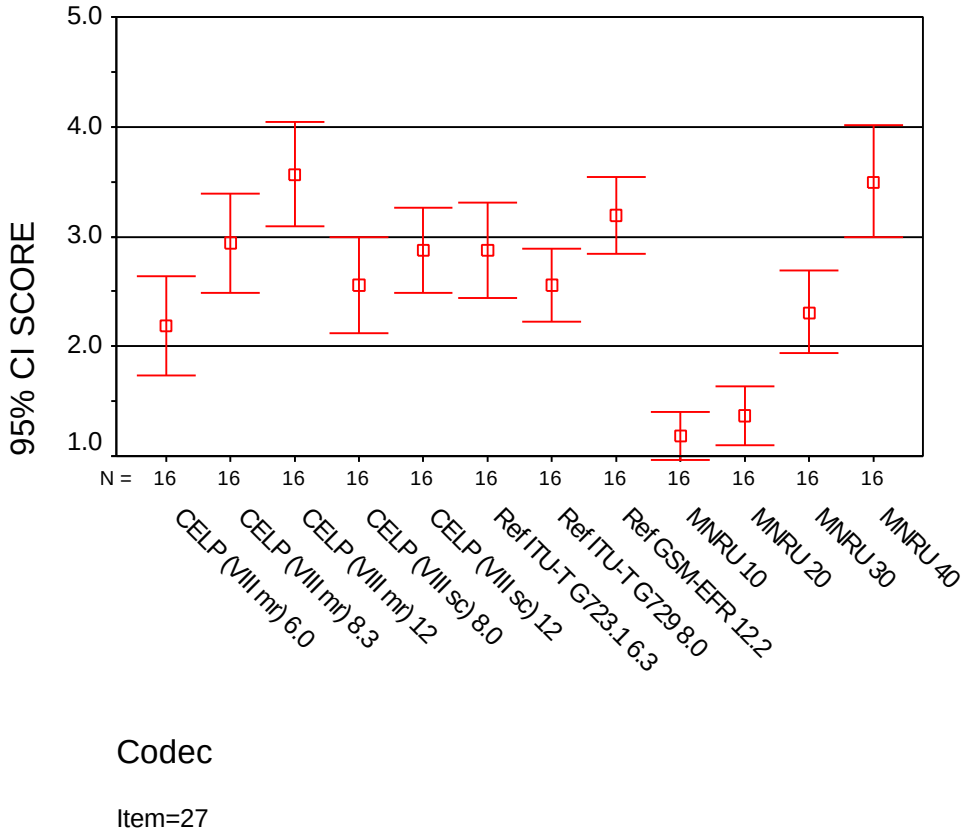
Item 136, Male (Swedish)



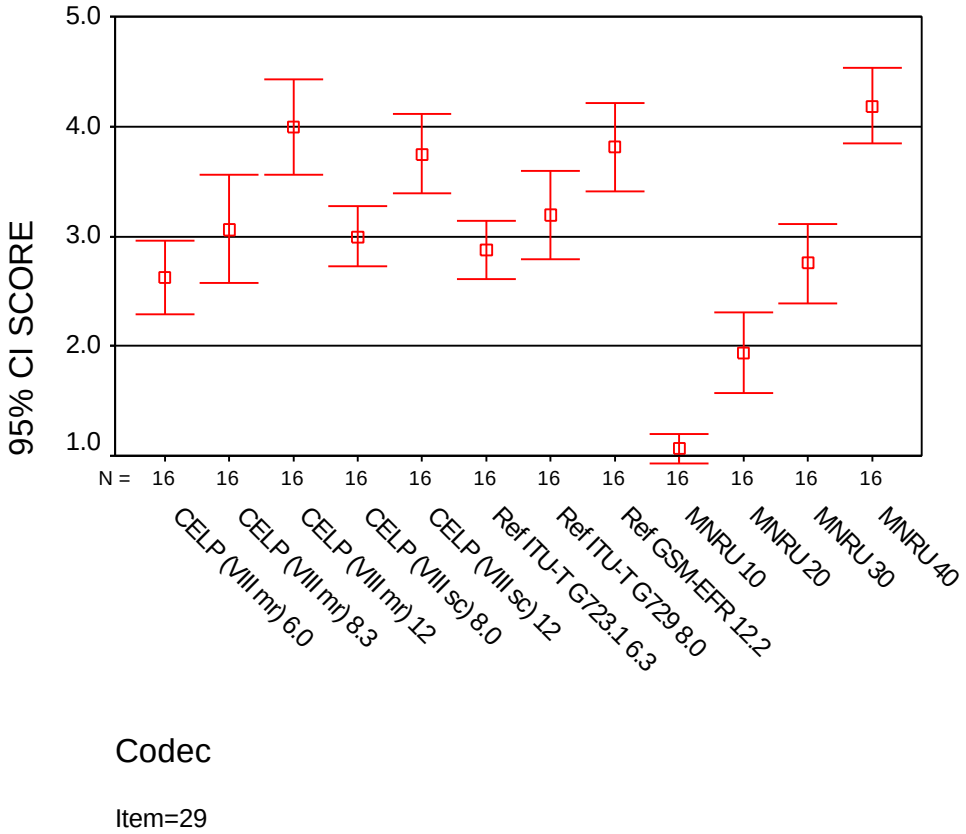
Item 26, Female (German)



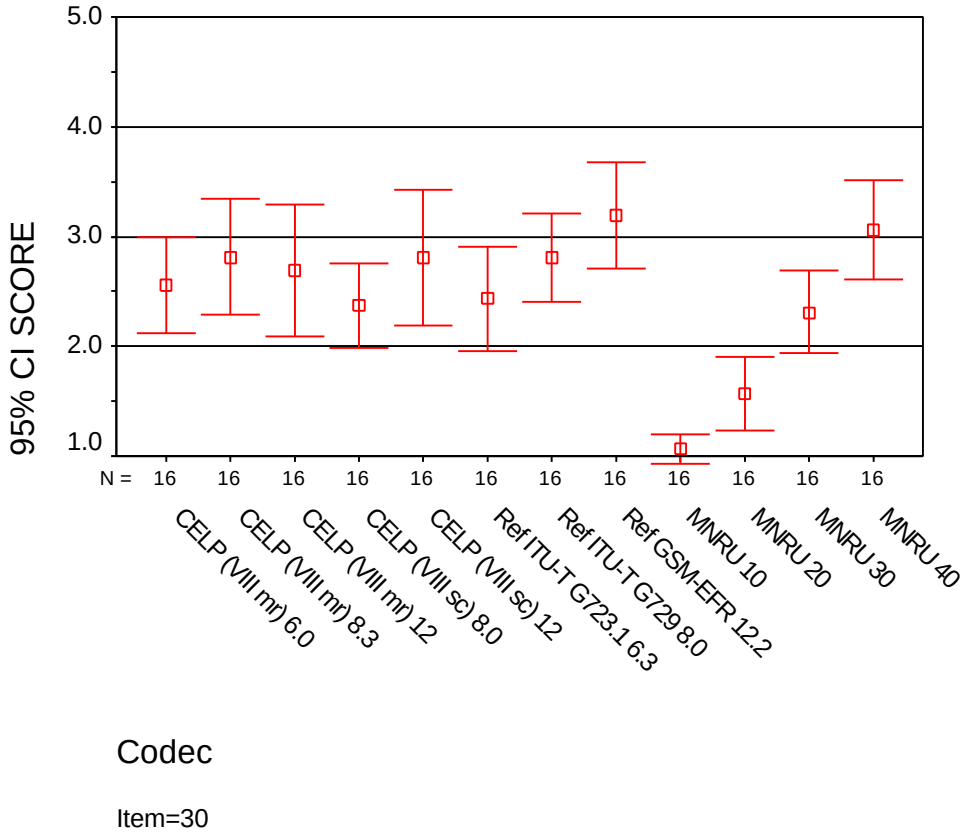
Item 27, Female (German)



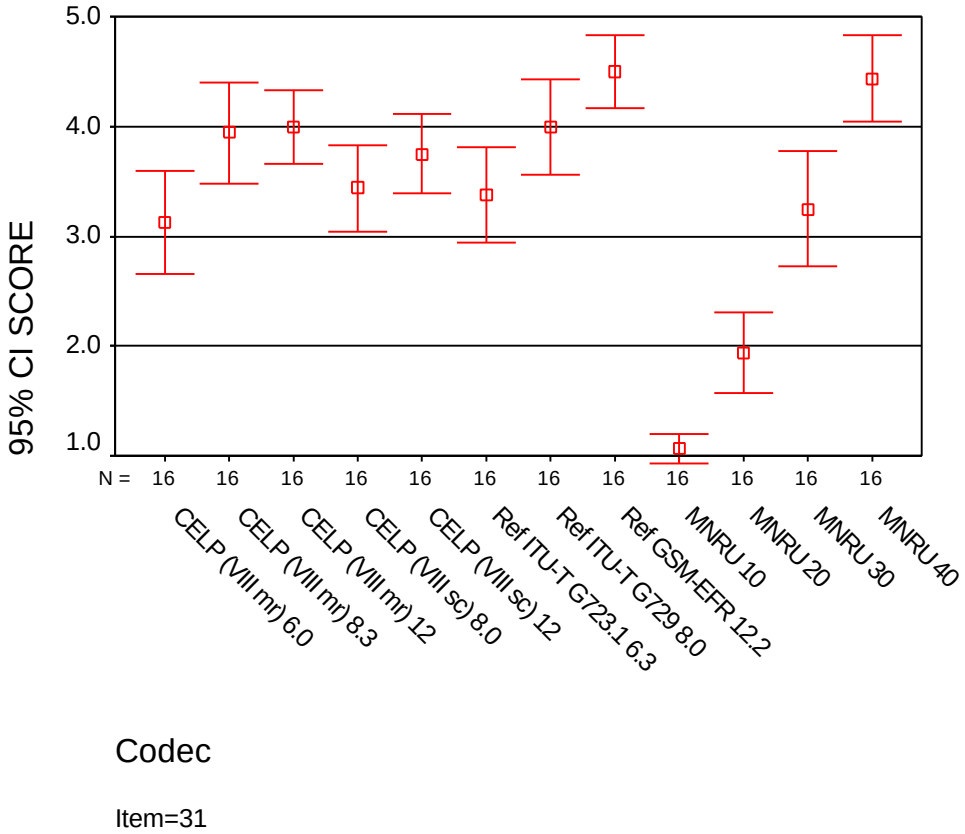
Item 29, Female (German)



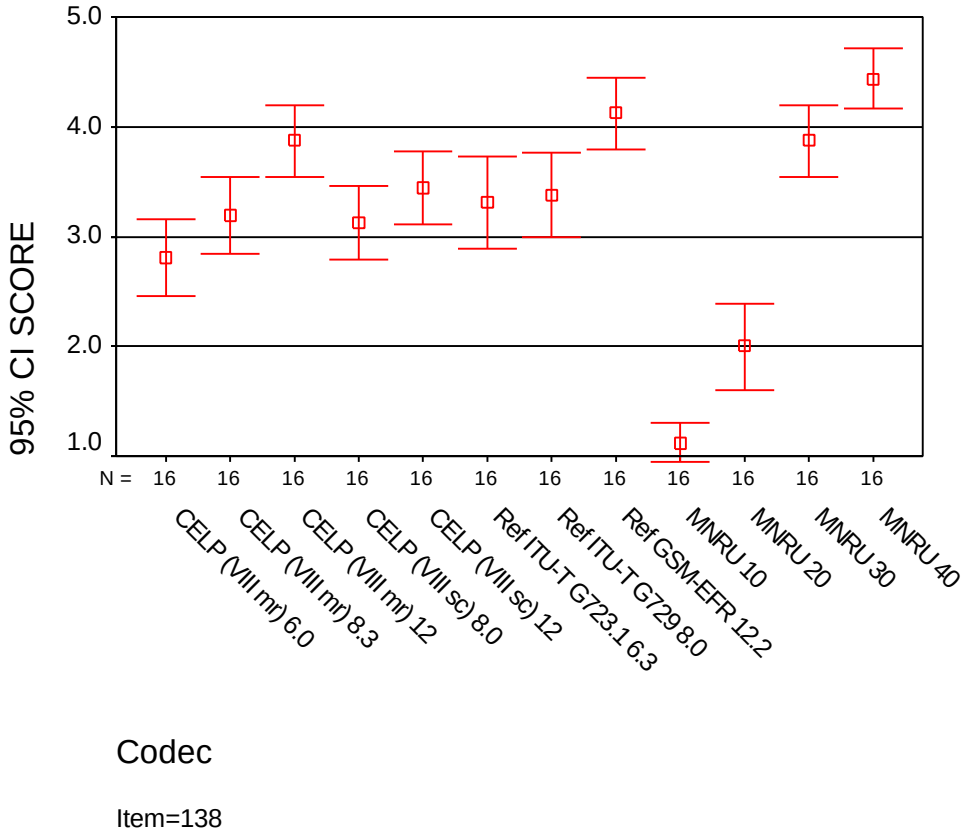
Item 30, Female (English)



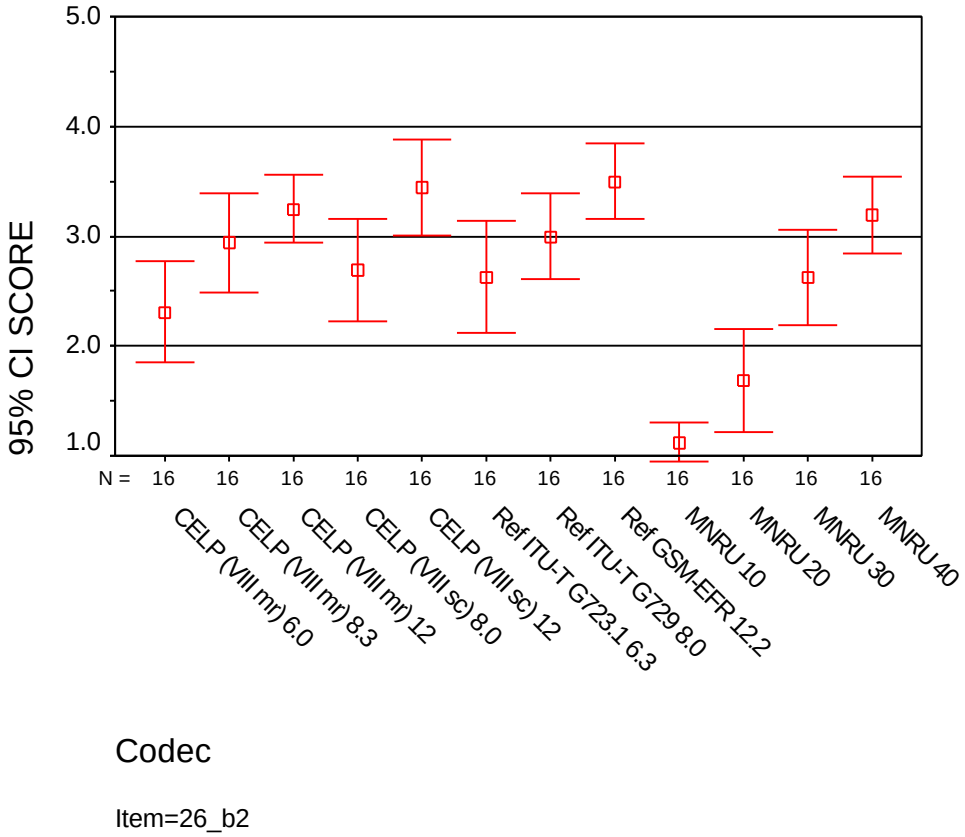
Item 31, Female (English)



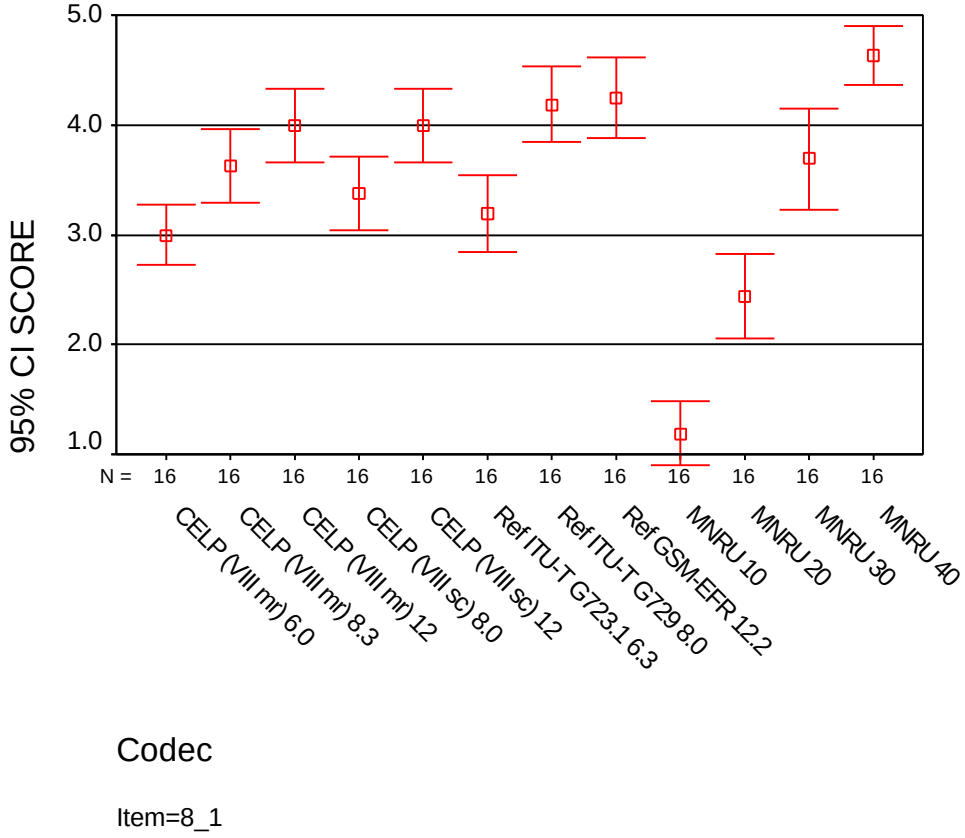
Item 138, Female (Swedish)



Item 26_b2, Female babble background noise (German)



Item 08_c1, Male Car background noise (English)



Item 55, Female background music (English)

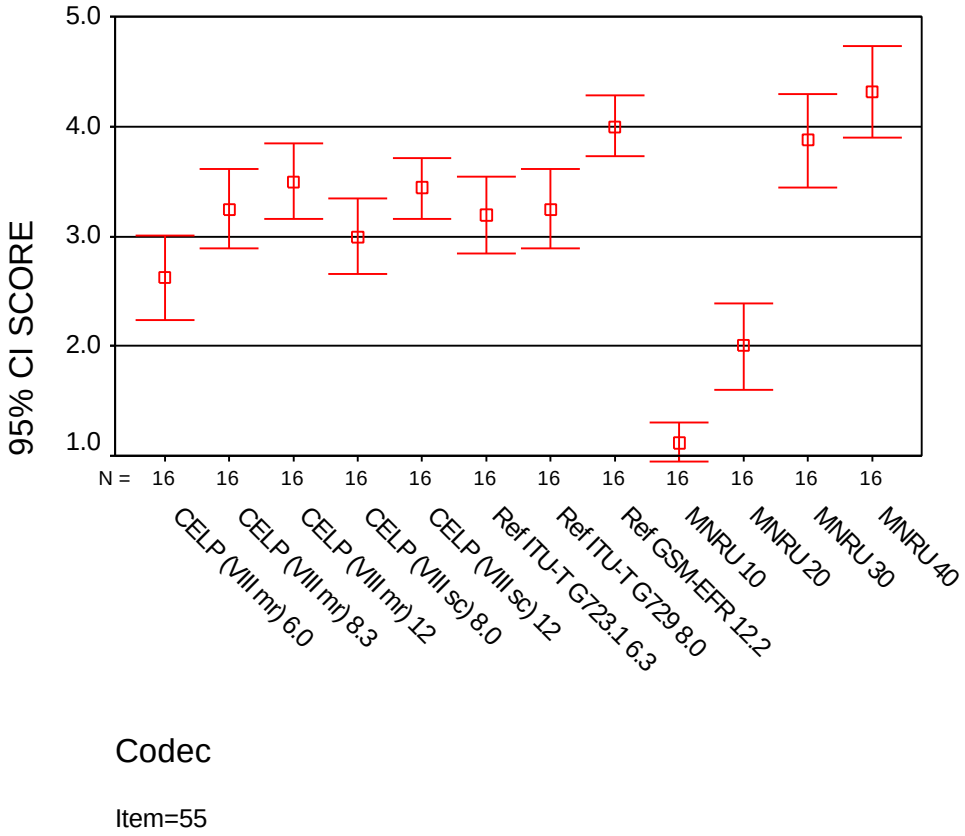
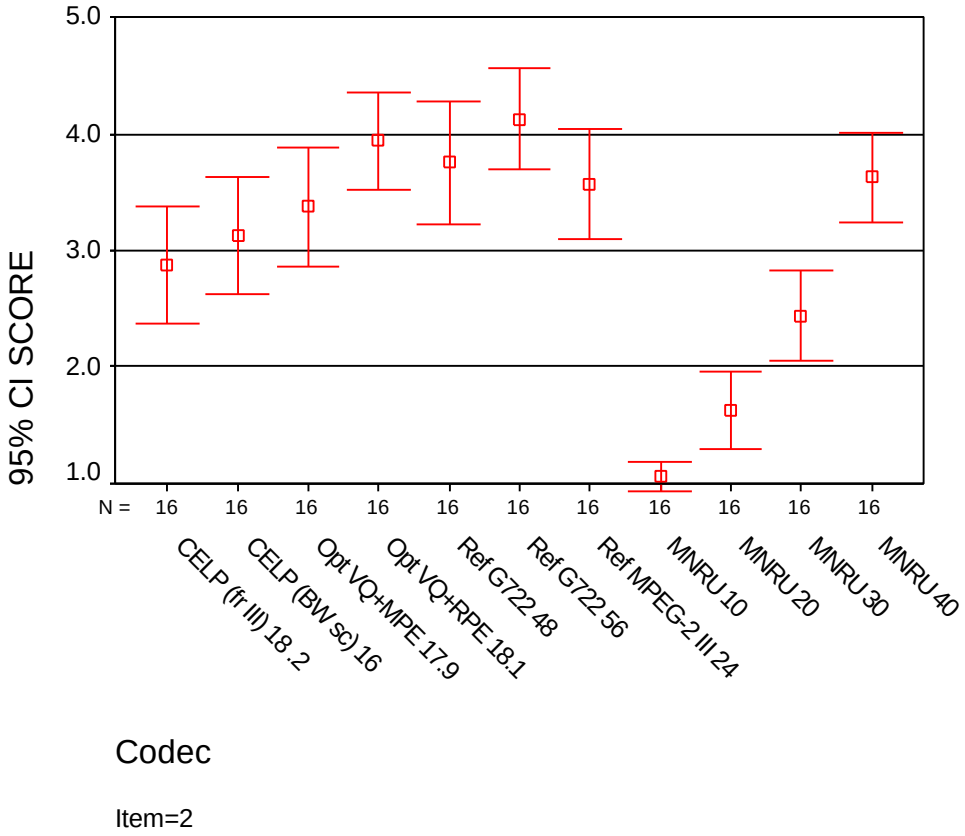
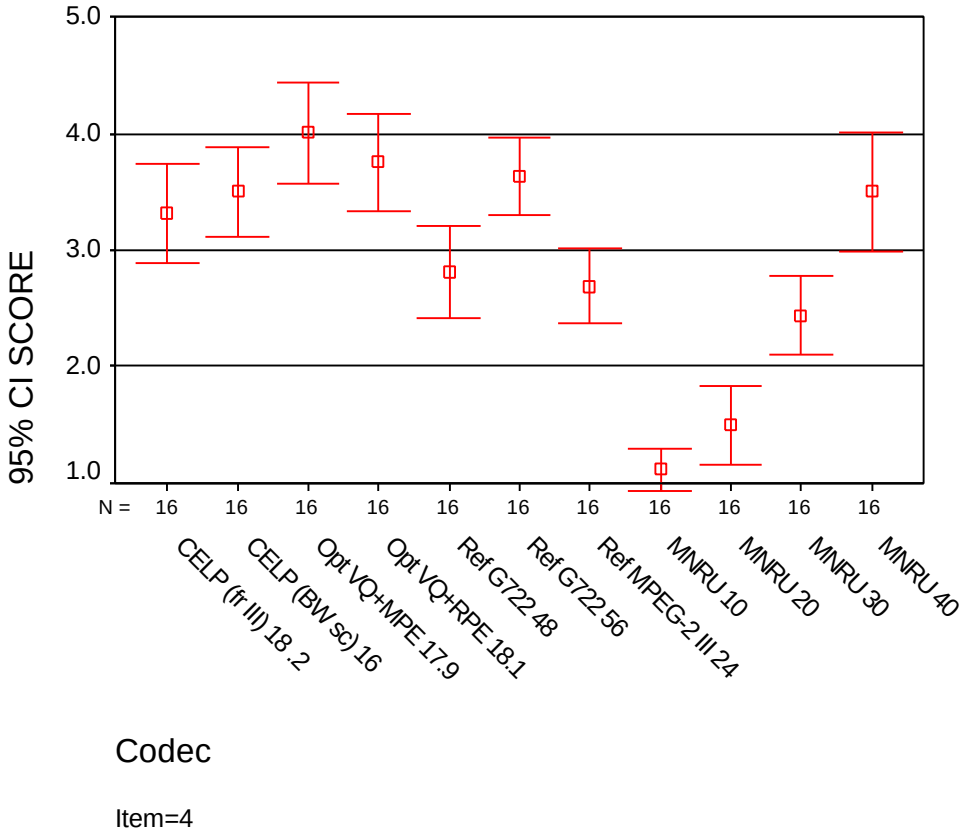


Figure 14. Results of the listening test 2 (NB-CELP).

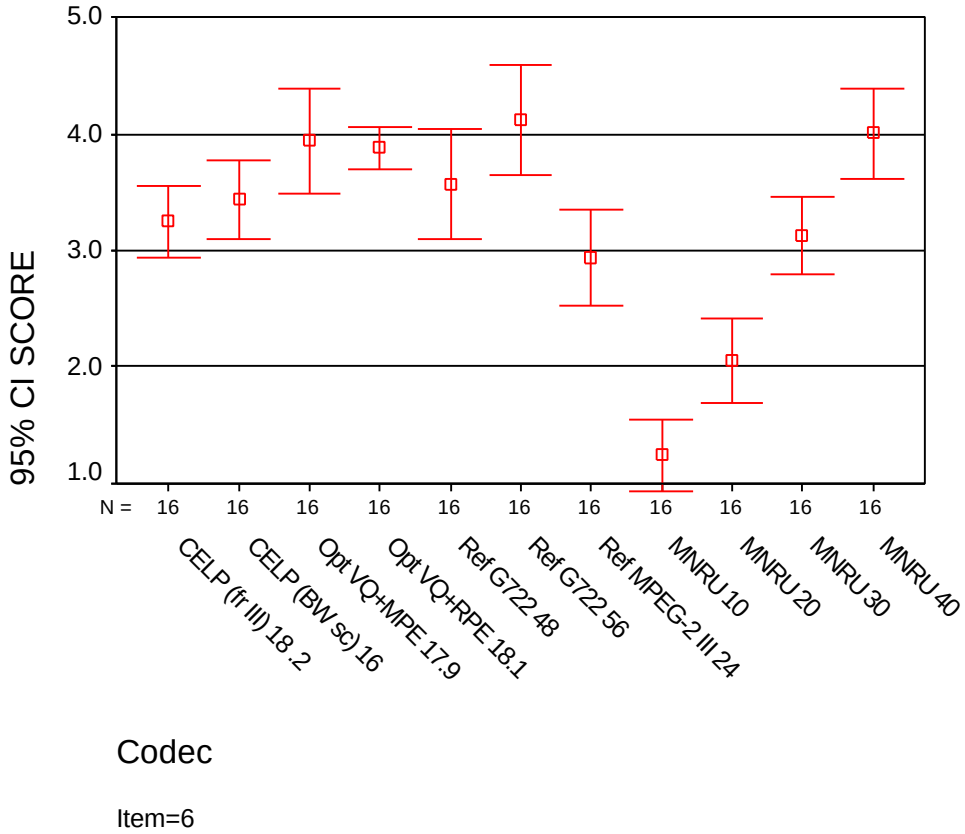
Item 02, Male (German)



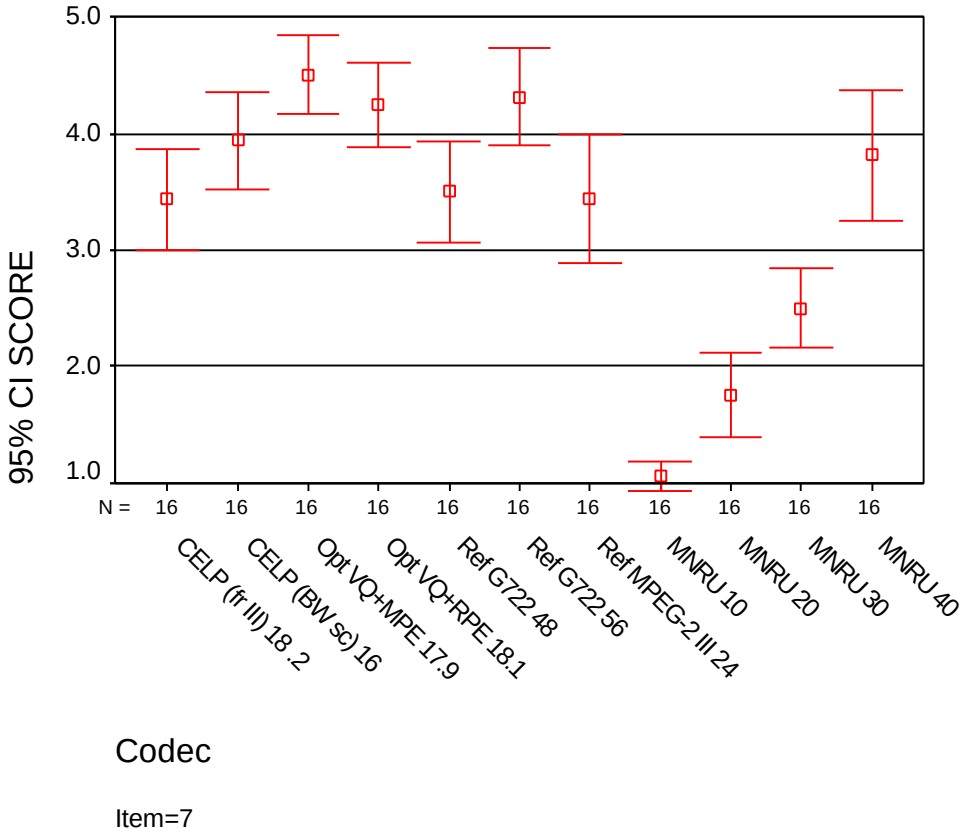
Item 04, Male (German)



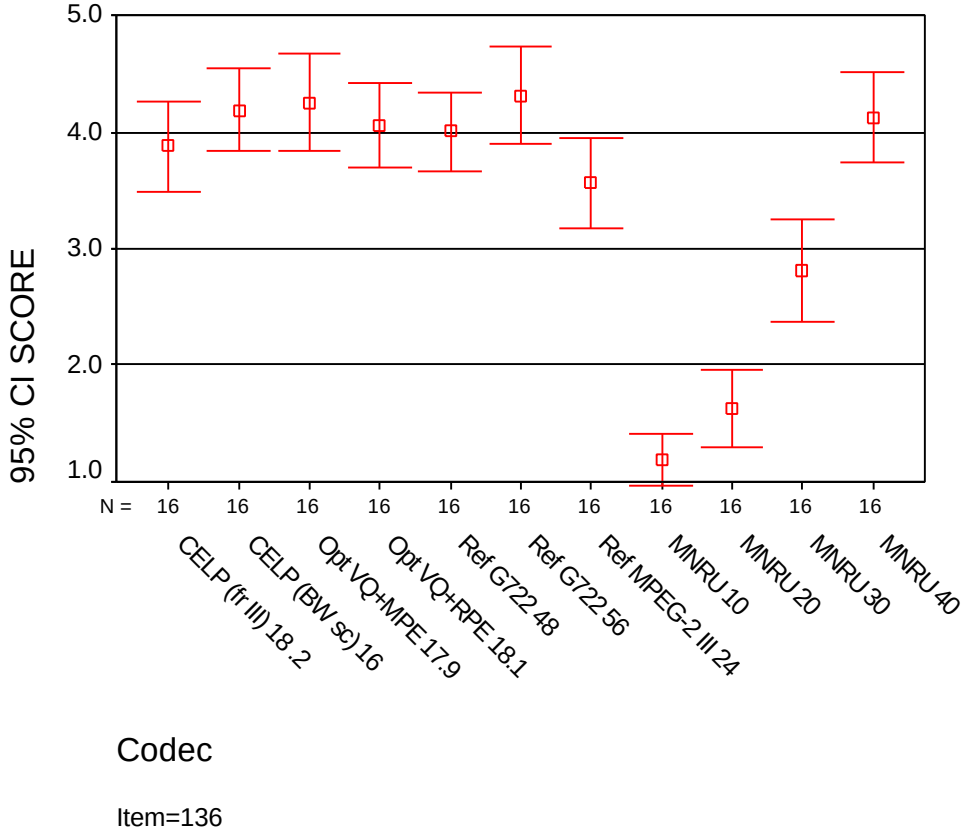
Item 06, Male (English)



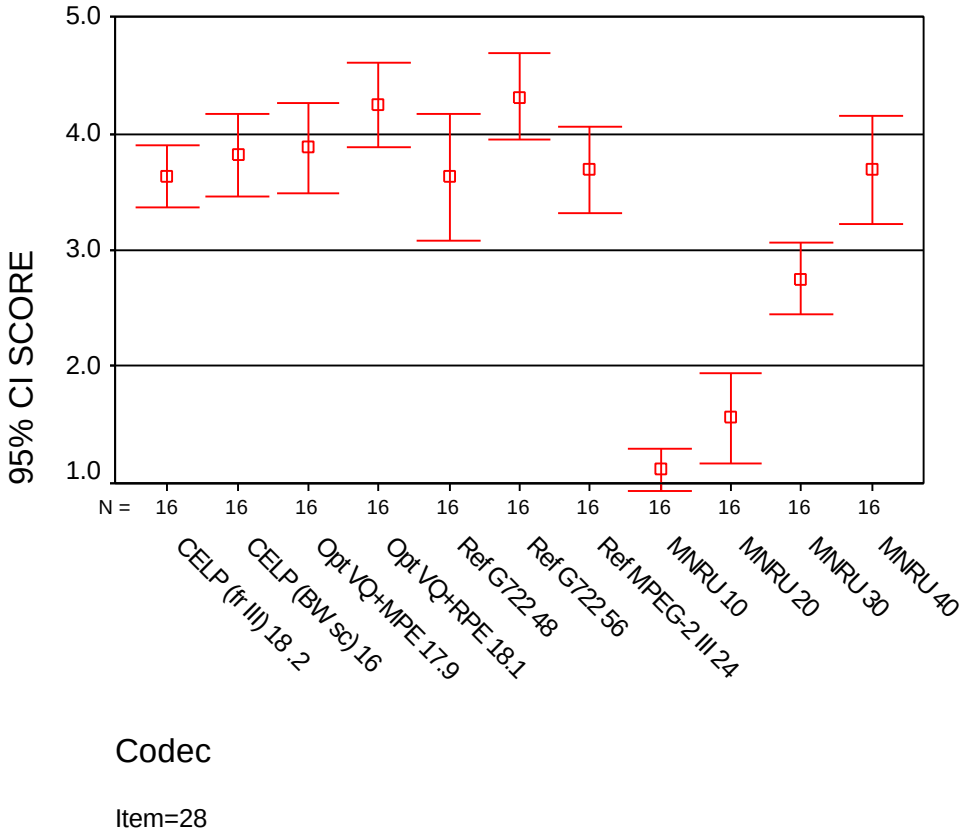
Item 07, Male (English)



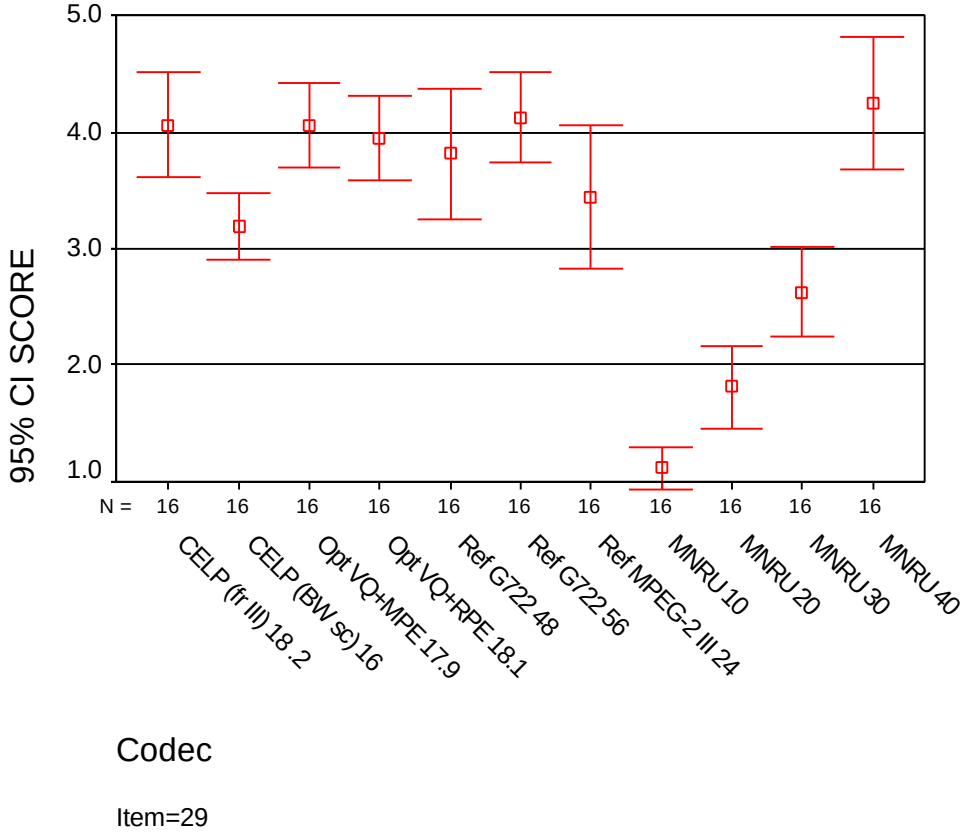
Item 136, Male (Swedish)



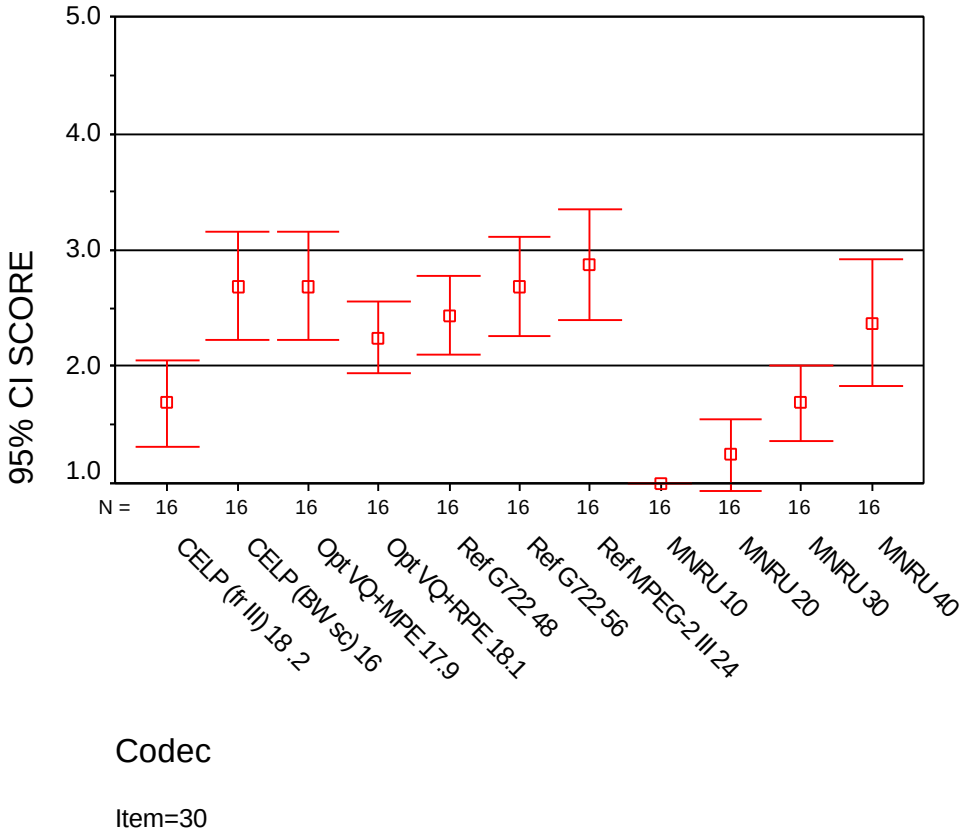
Item 28, Female (German)



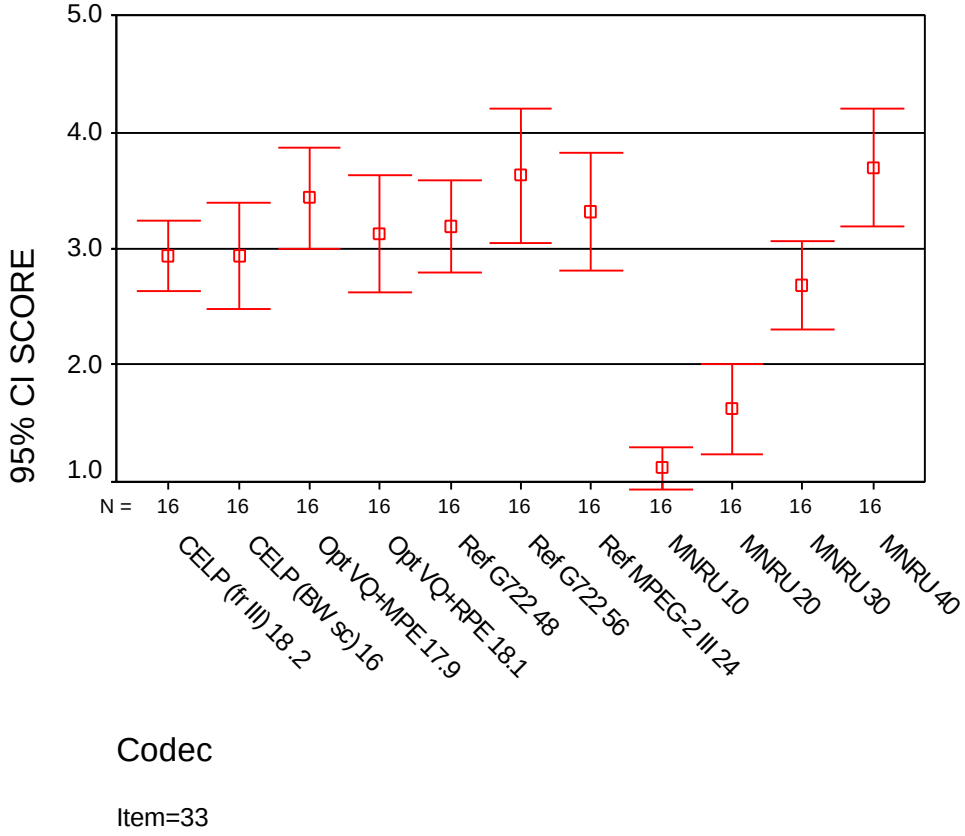
Item 29, Female (German)



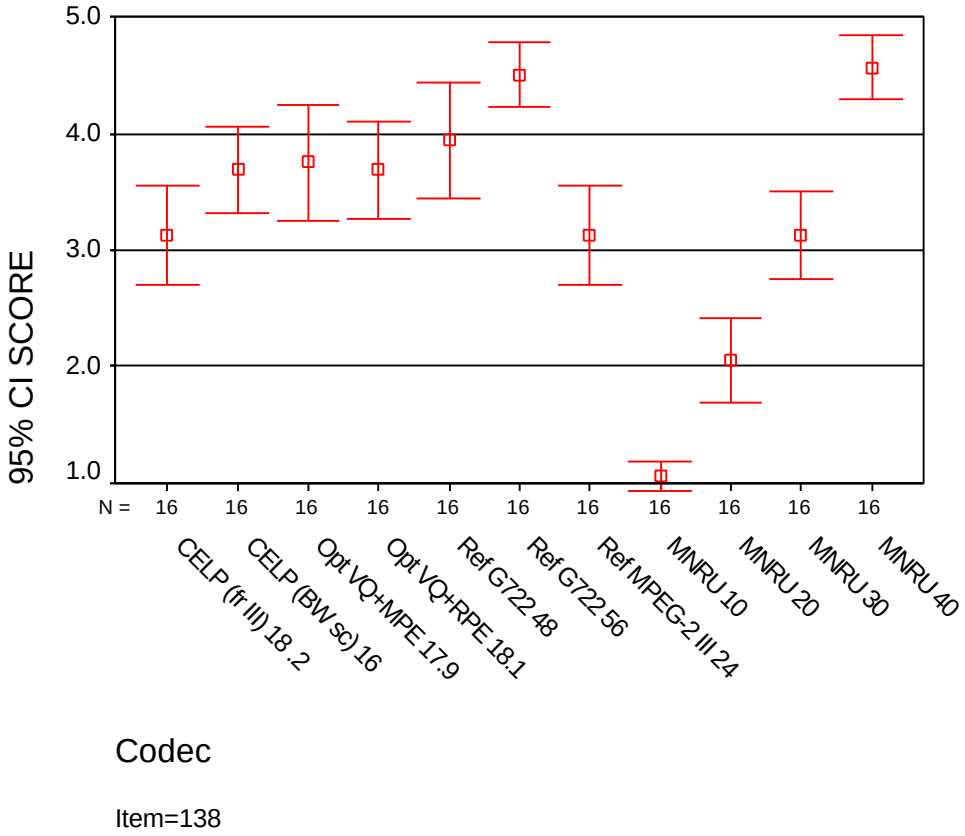
Item 30, Female (English)



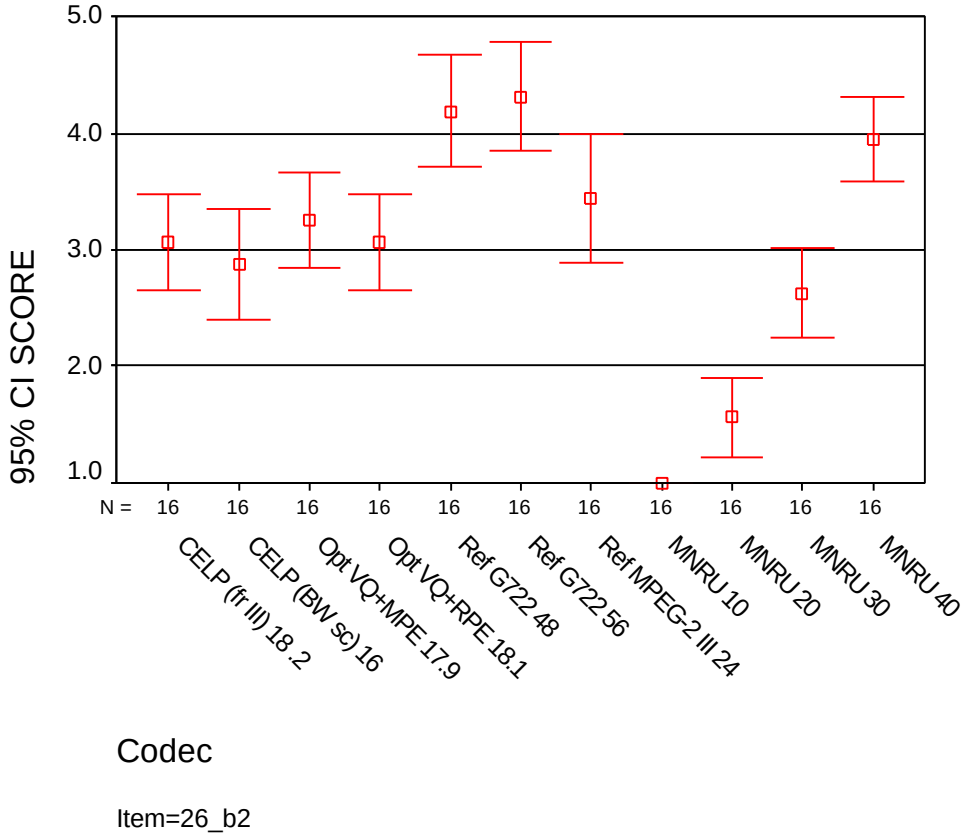
Item 33, Female (English)



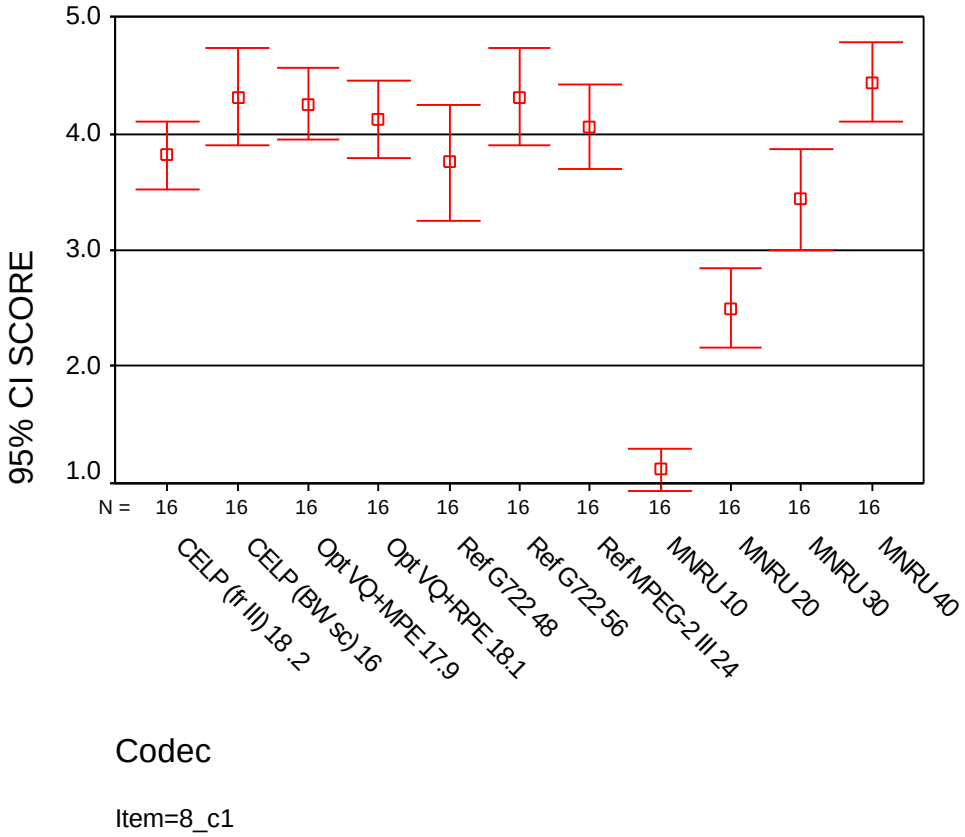
Item 138, Female (Swedish)



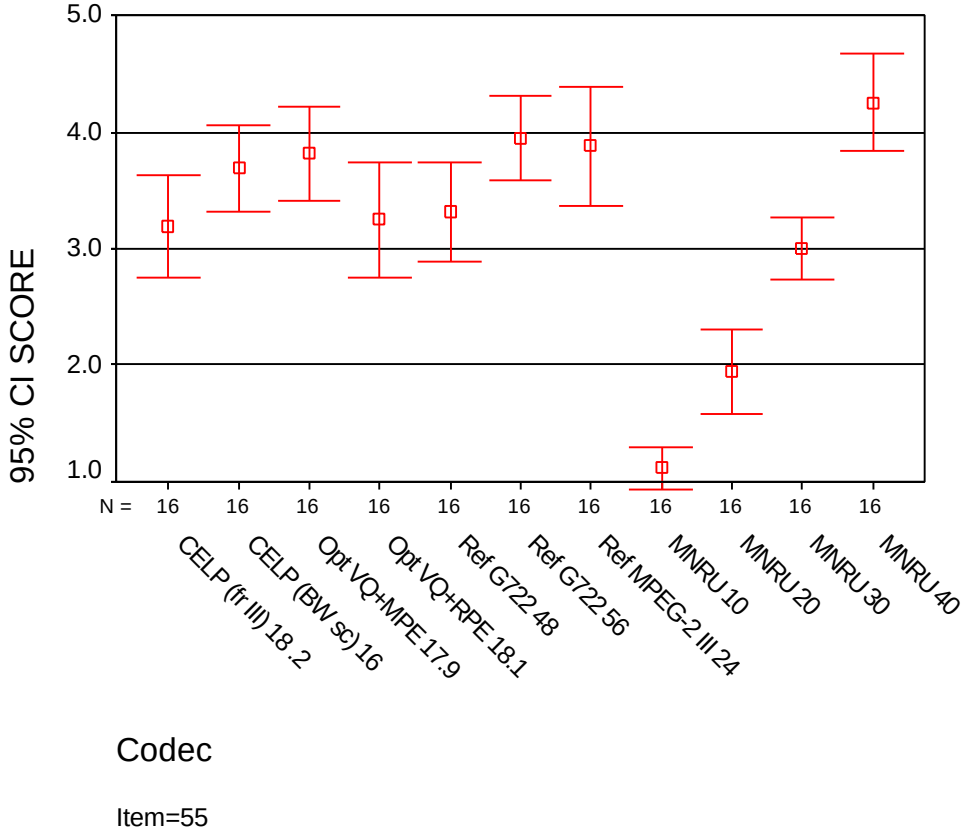
Item 26_b2, Female babble background noise (German)



Item 08_c1, Male Car background noise (English)



Item 55, Female background music (English)



Item 83, Classical music

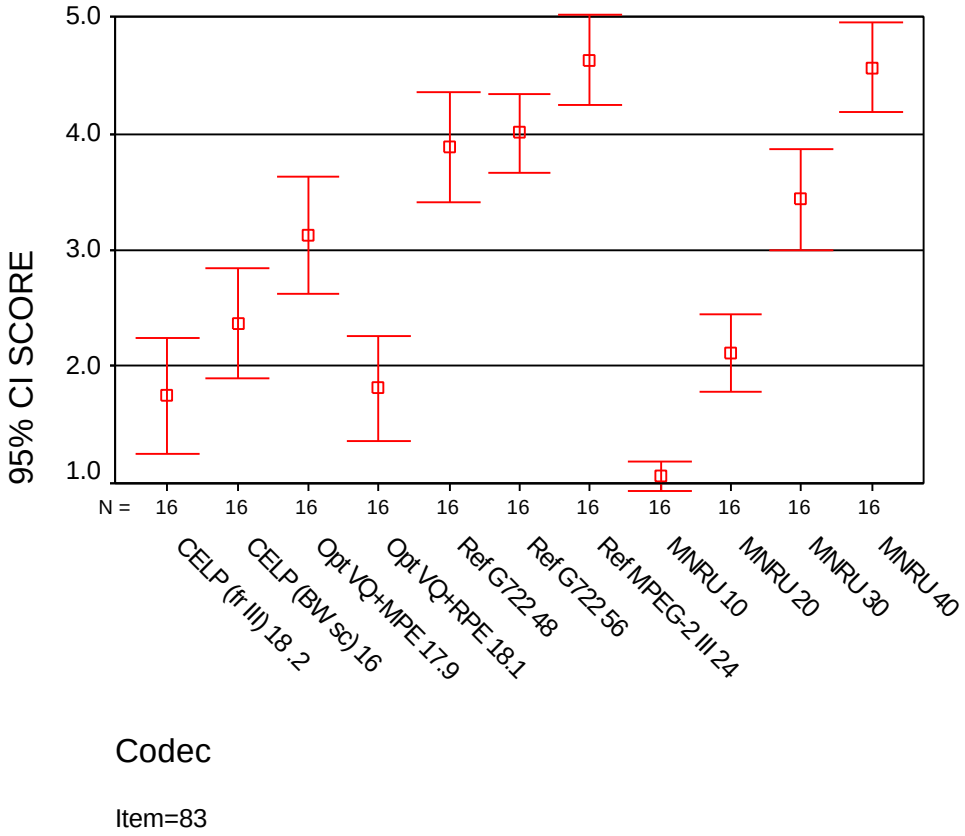
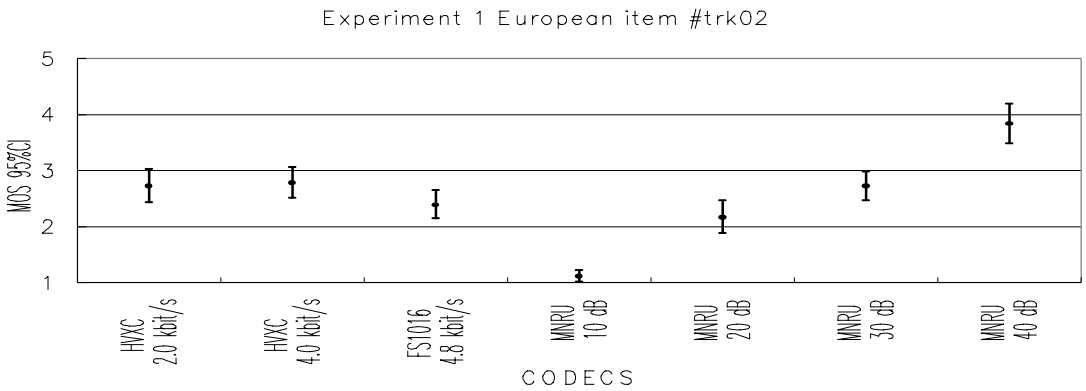


Figure 15. Item by item results of the listening test 3 (WB-CELP).

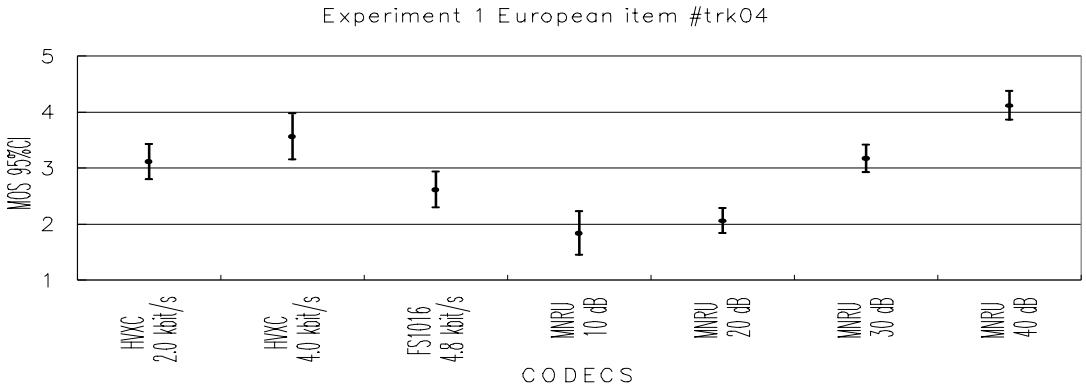
9.3.2 FhG site

The performance of each coder in experiments 1, 2 and 3 are shown graphically in Figures 16, 17 and 18, respectively. In Figure 15, CELP 8.0 kbit/s should be written as CELP 8.3 kbit/s.

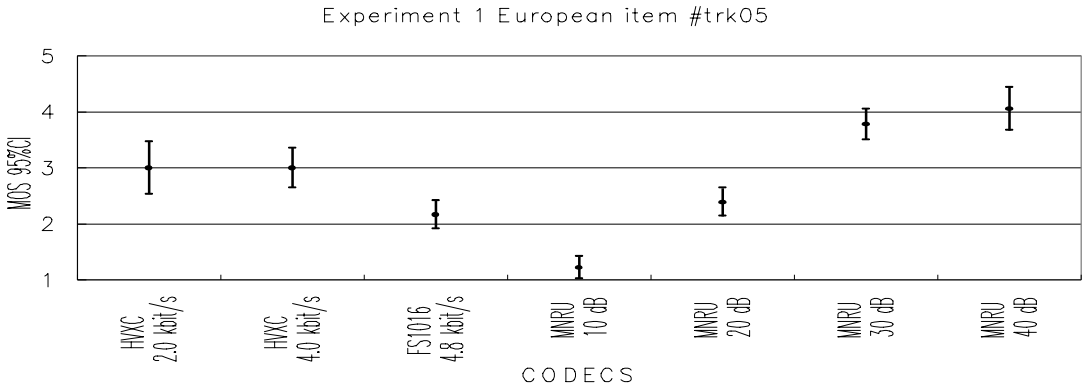
Item 02, Male (German)



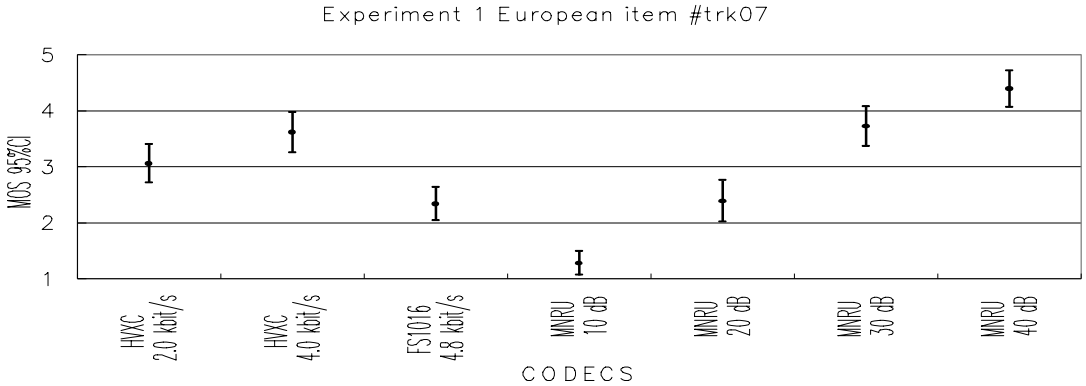
Item 04, Male (German)



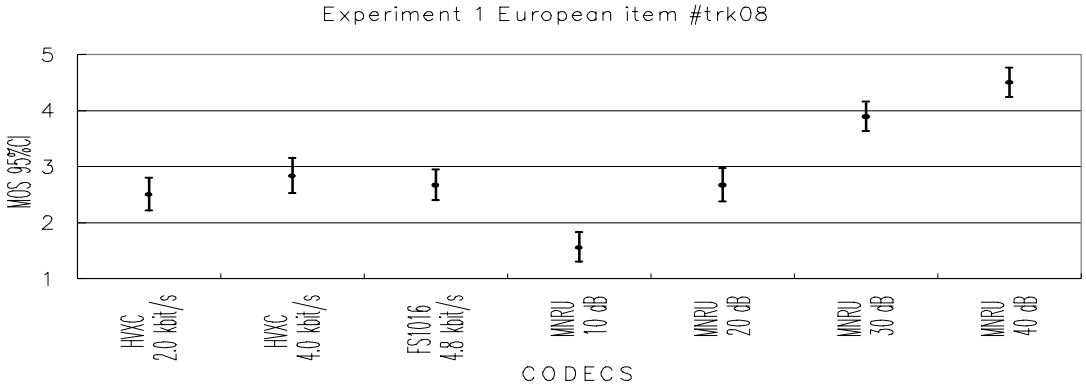
Item 05, Male (German)



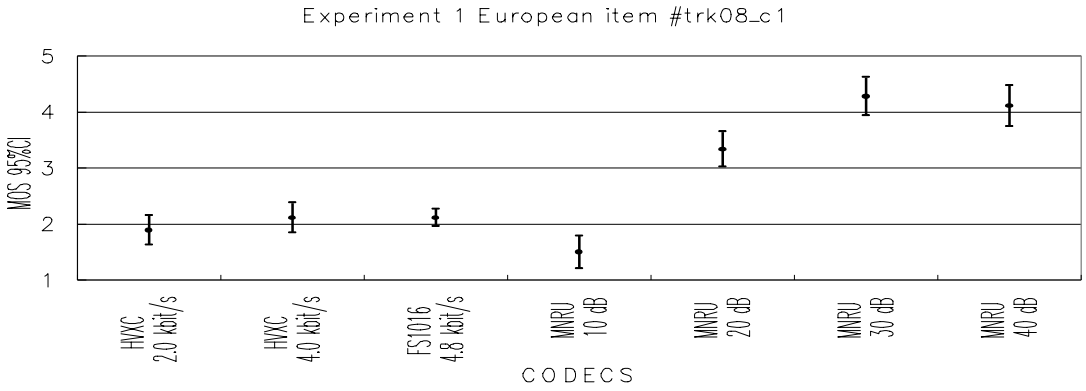
Item 07, Male (English)



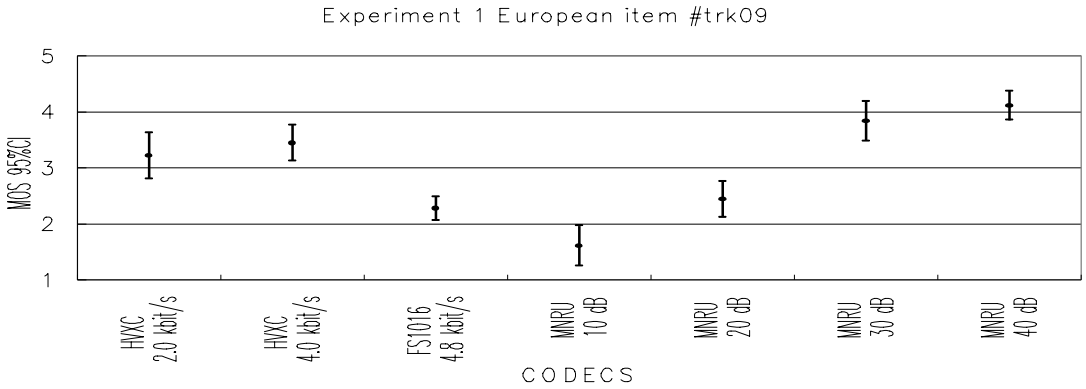
Item 08, Male (English)



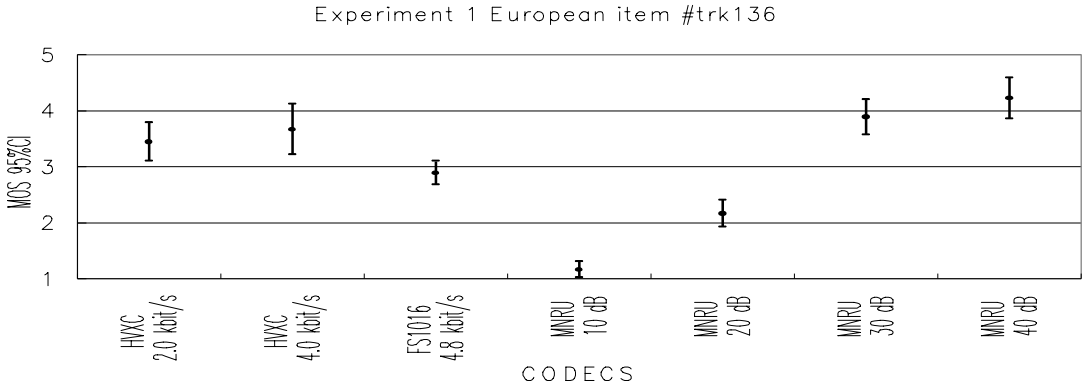
Item 08_c1, Male with car background noise (English)



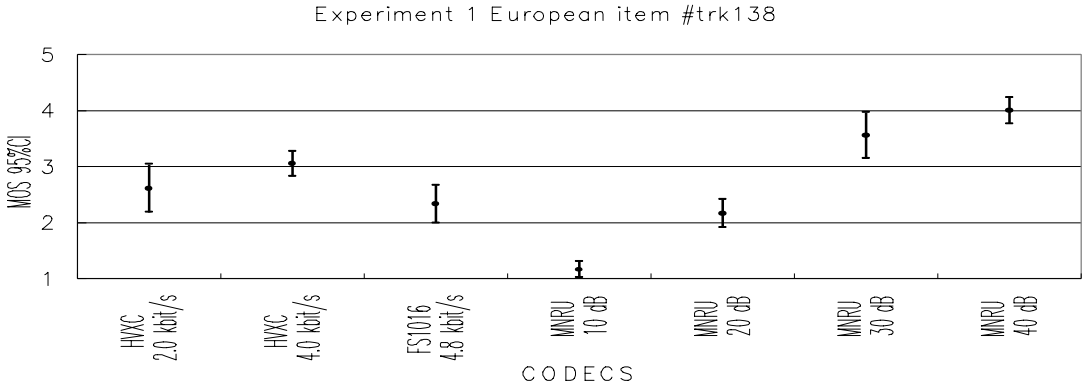
Item 09, Male (English)



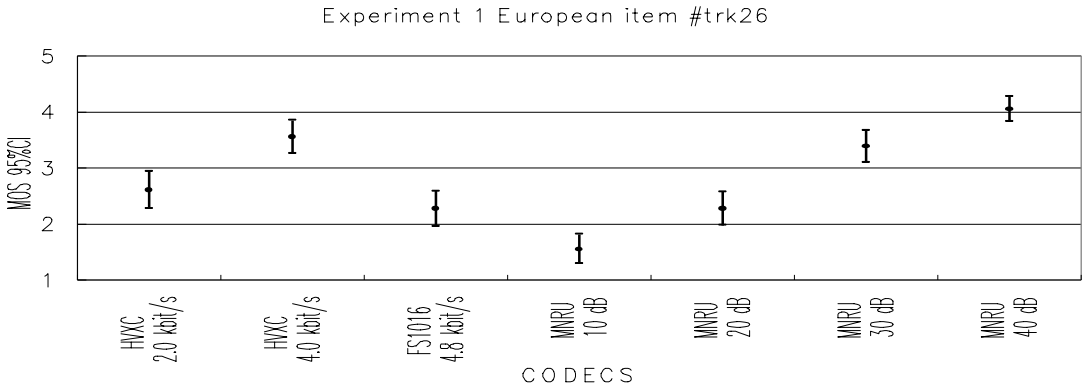
Item 136, Male (Swedish)



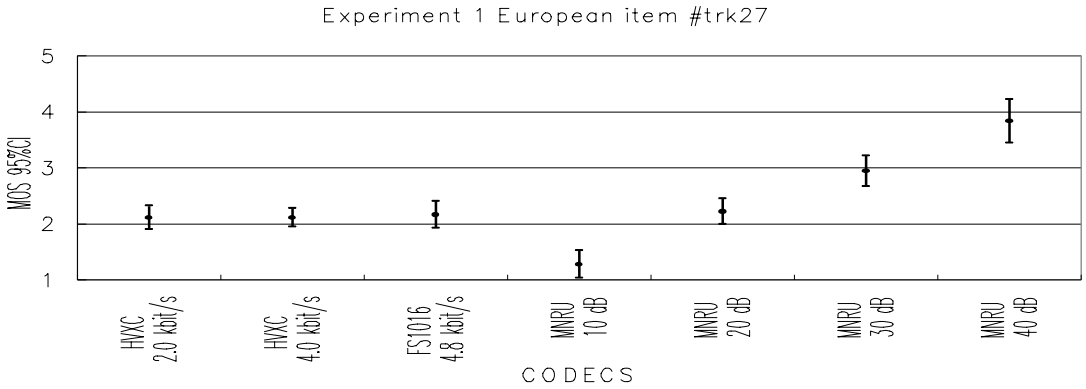
Item 138, Female (Swedish)



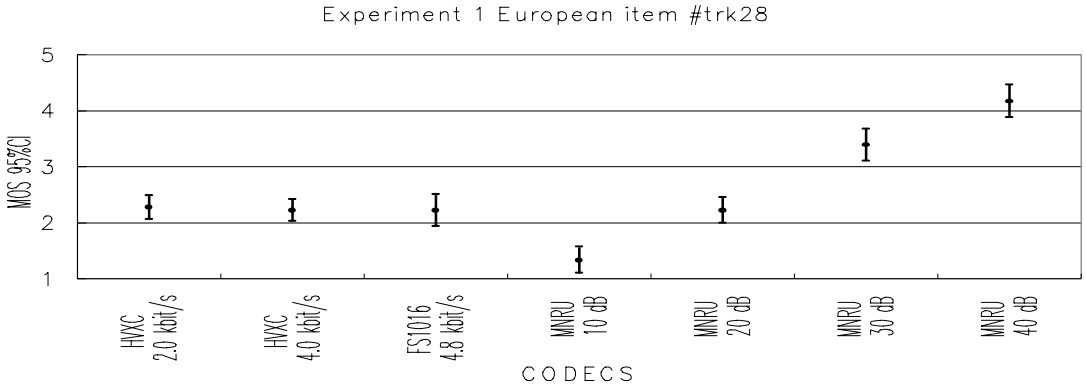
Item 26, Female (German)



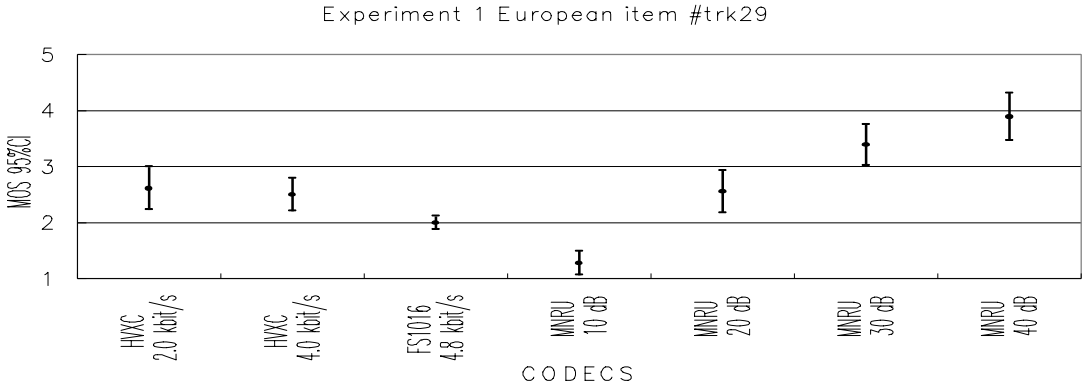
Item 27, Female (German)



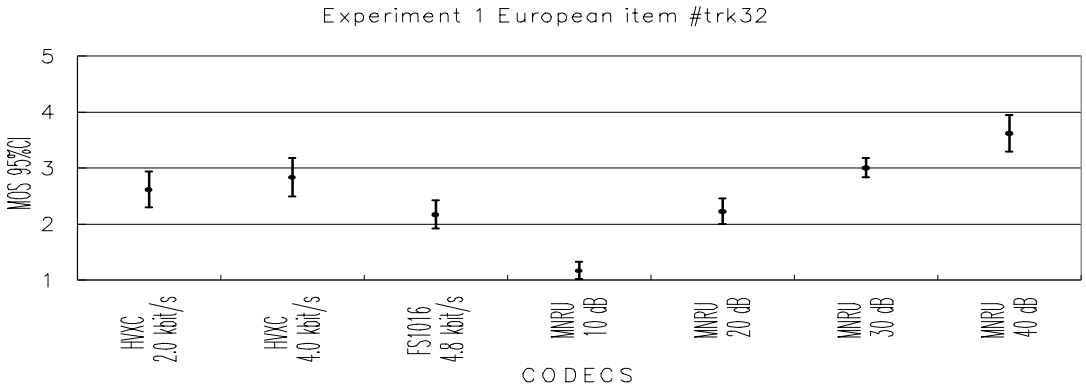
Item 28, Female (German)



Item 29, Female (German)



Item 32, Female (English)



Item 33, Female (English)

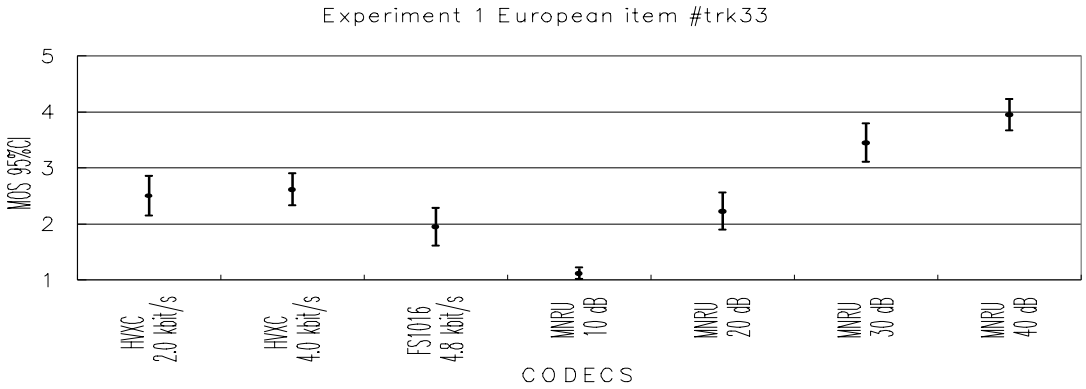
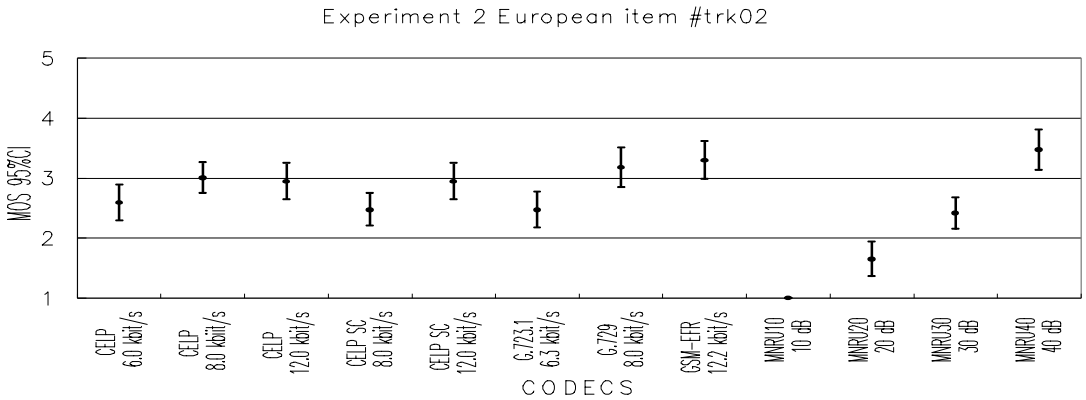
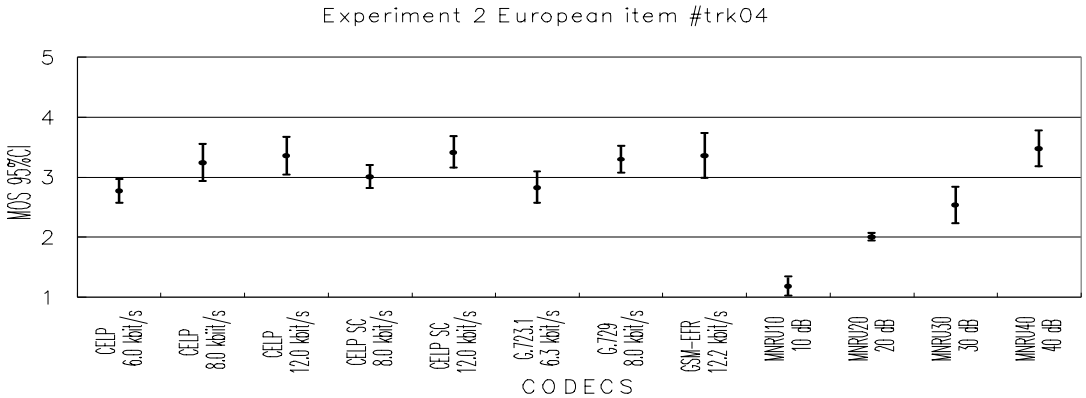


Figure 16. Item by item results of the listening test 1 (PARAMETRIC).

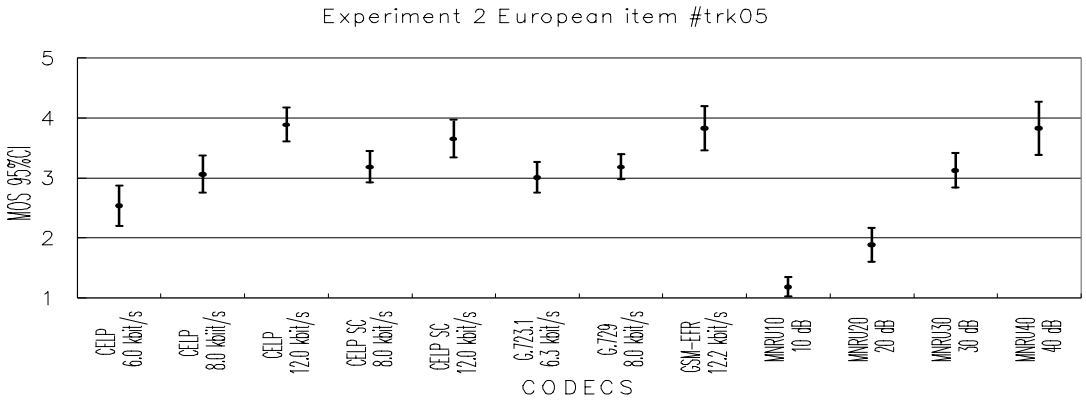
Item 02, Male (German)



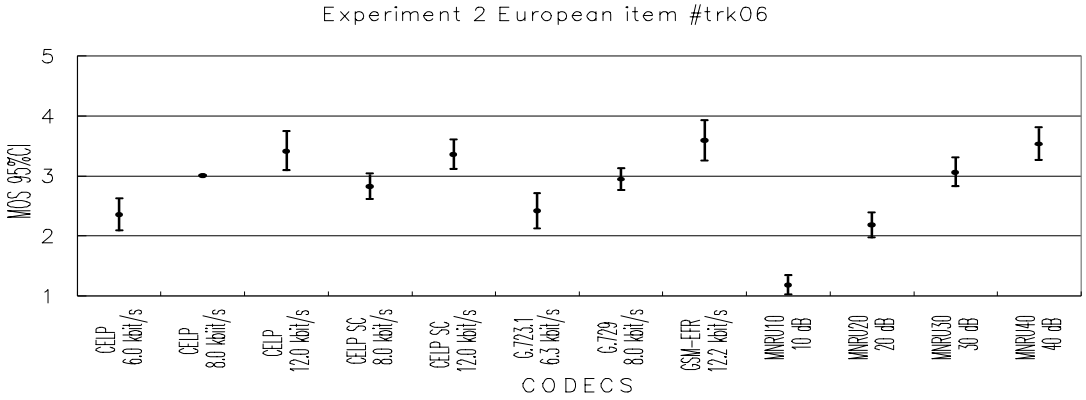
Item 04, Male (German)



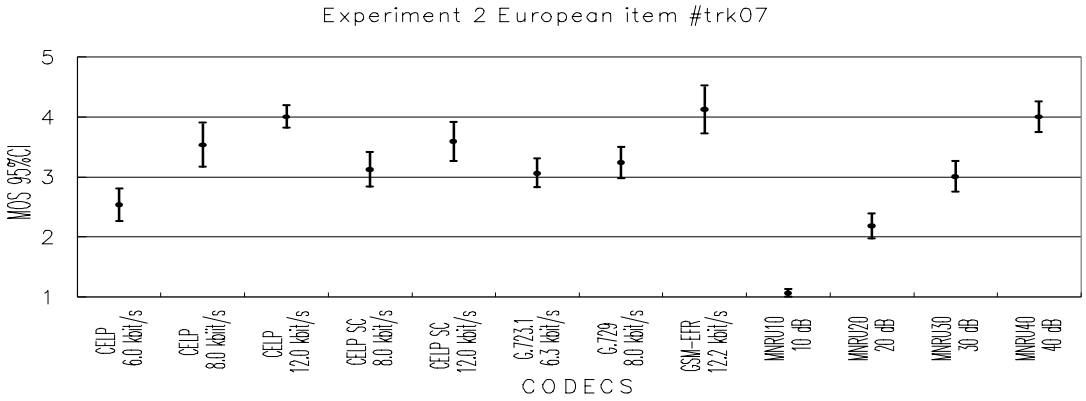
Item 05, Male (German)



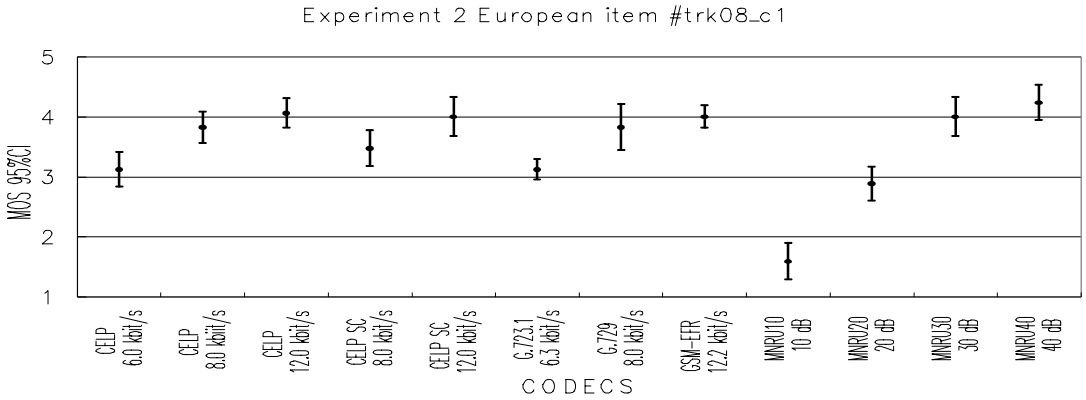
Item 06, Male (English)



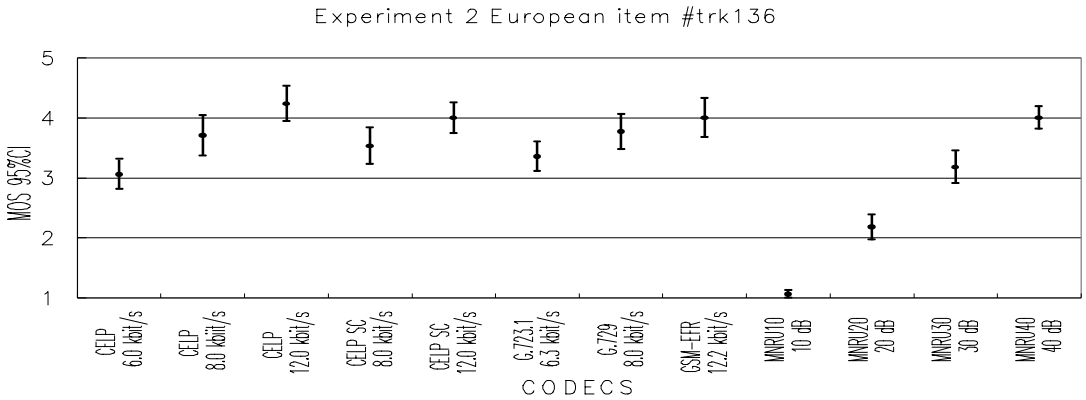
Item 07, Male (English)



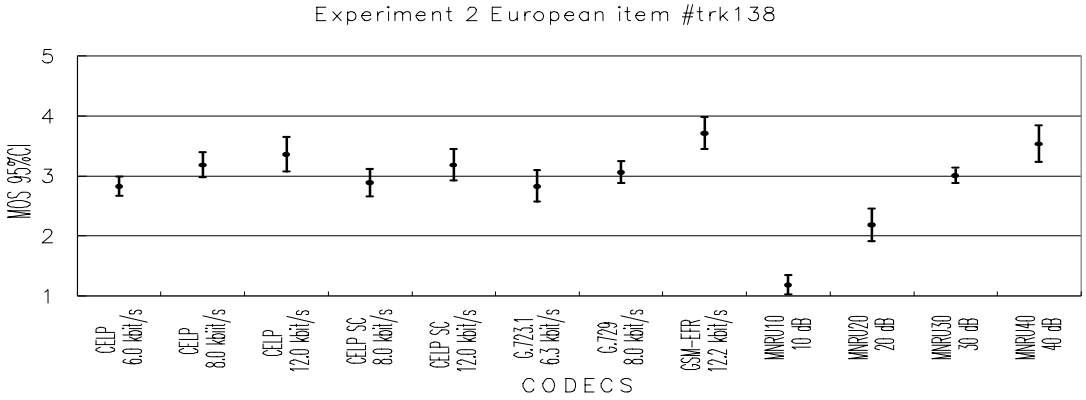
Item 08_c1, Male with car background noise (English)



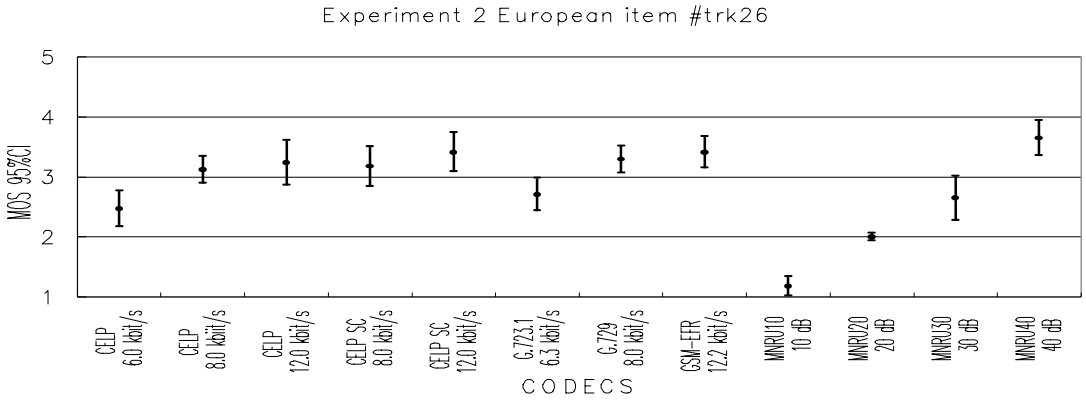
Item 136, Male (Swedish)



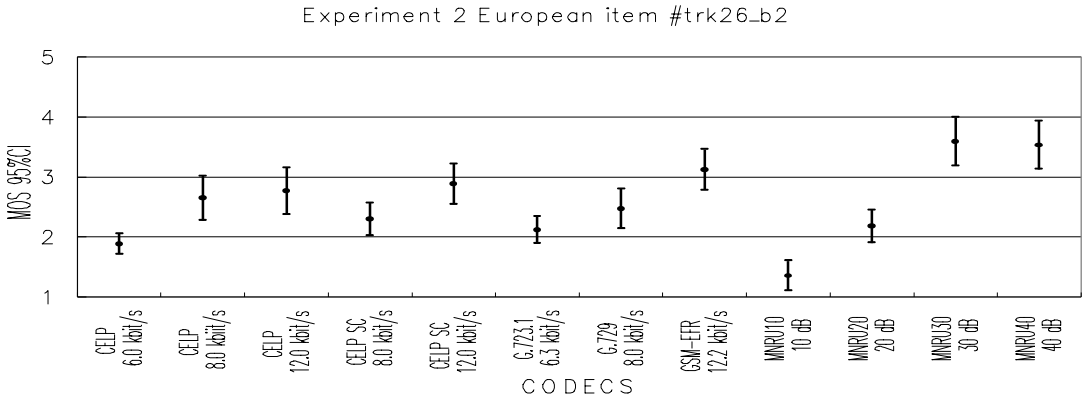
Item 138, Male (Swedish)



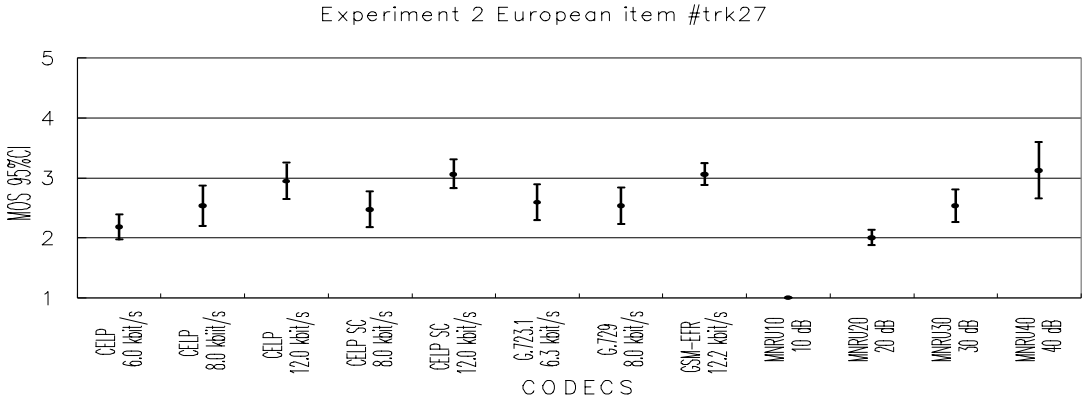
Item 26, Female (German)



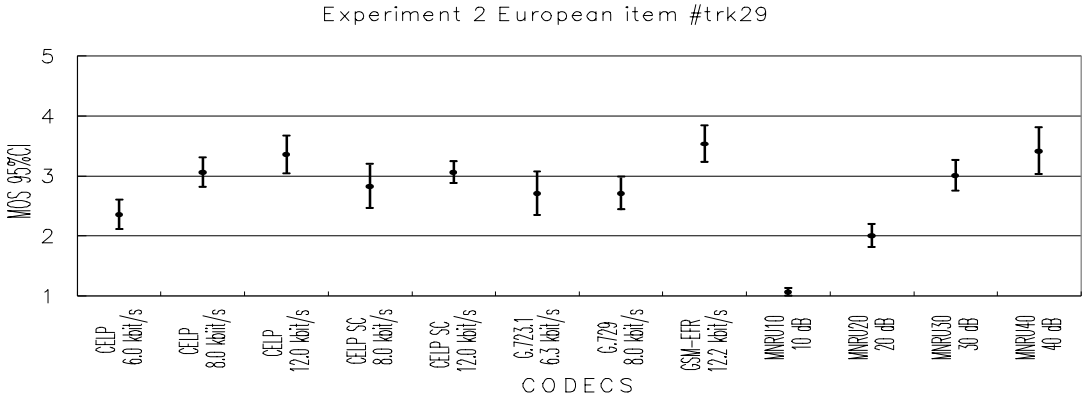
Item 26_b2, Female with babble background noise (German)



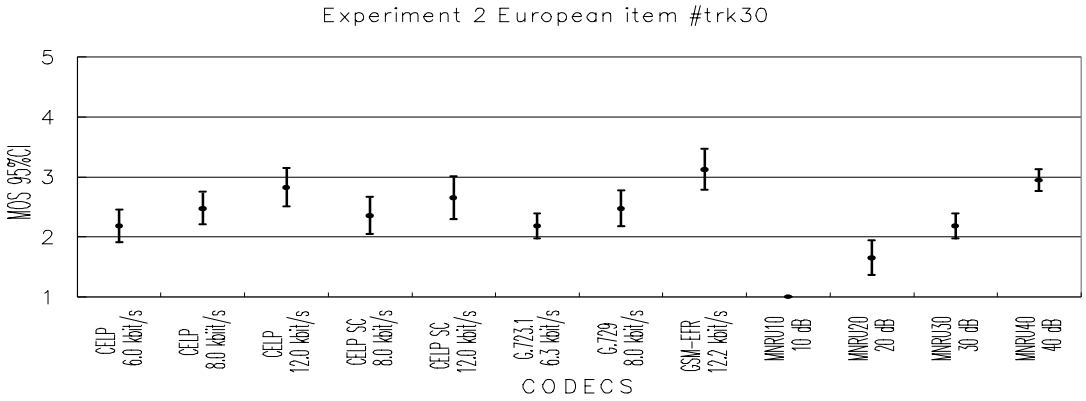
Item 27, Female (German)



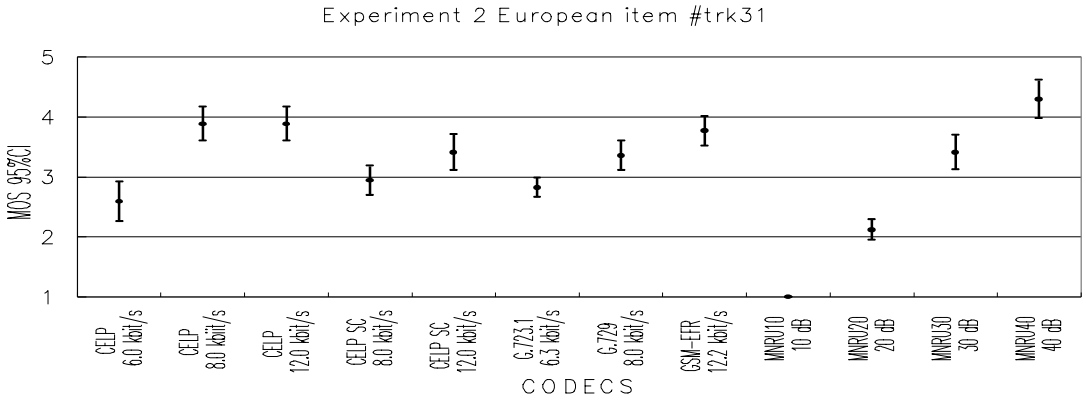
Item 29, Female (German)



Item 30, Female (English)



Item 31, Female (English)



Item 55, Background music (English)

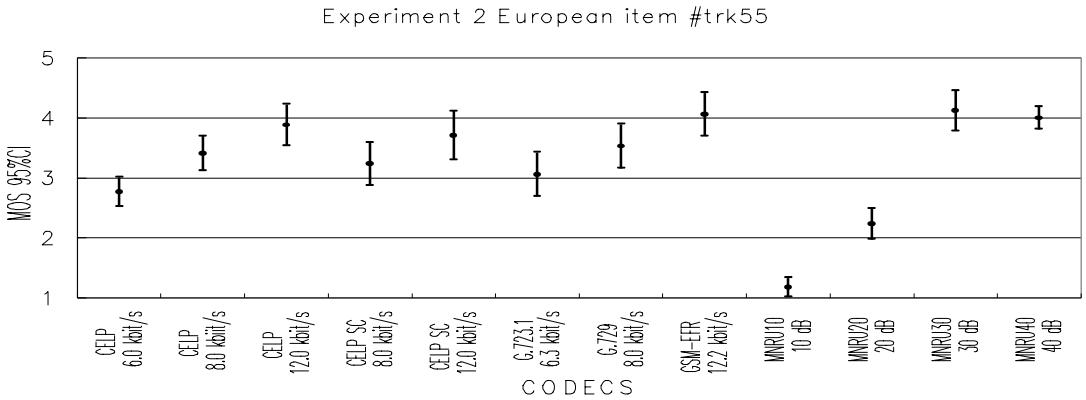
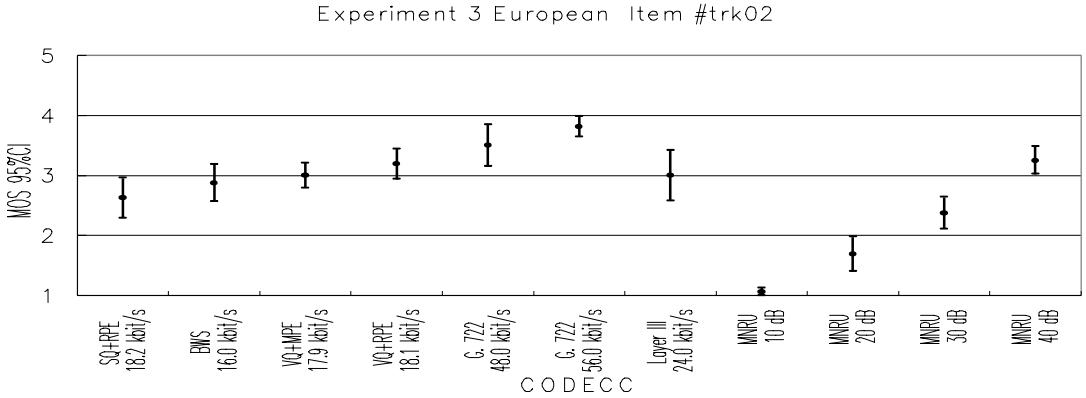
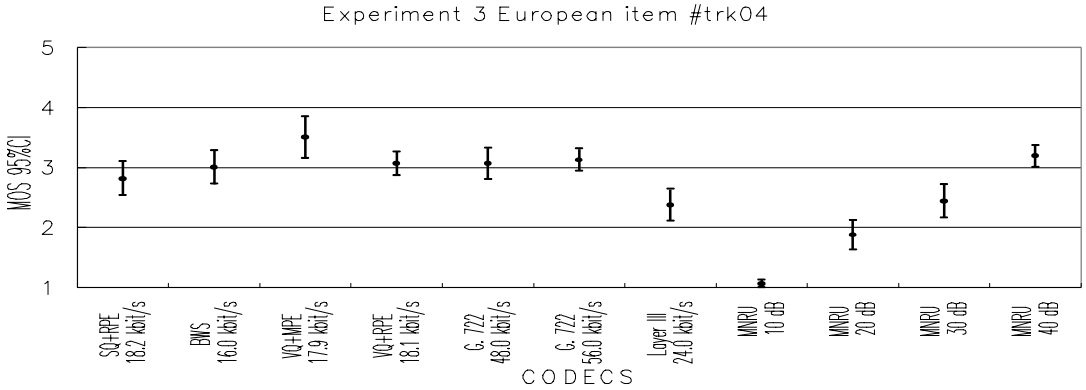


Figure 17. Item by item results of the listening test 2 (NB-CELP).

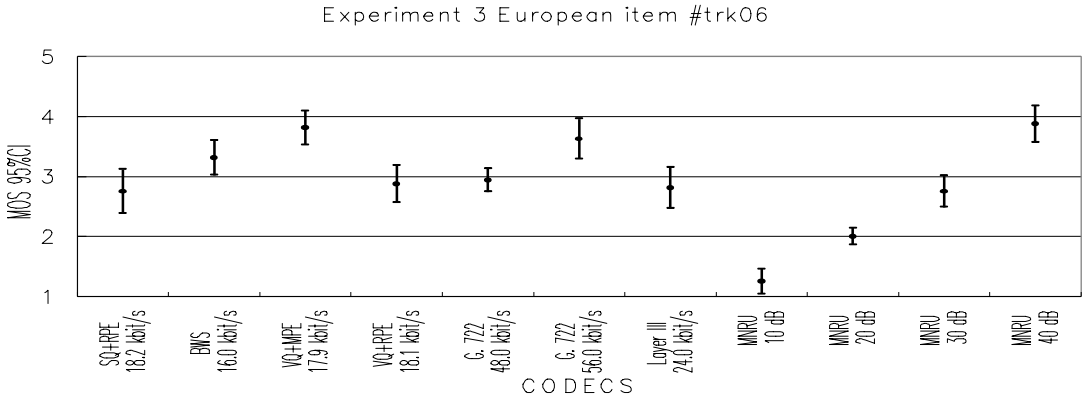
Item 02, Male (German)



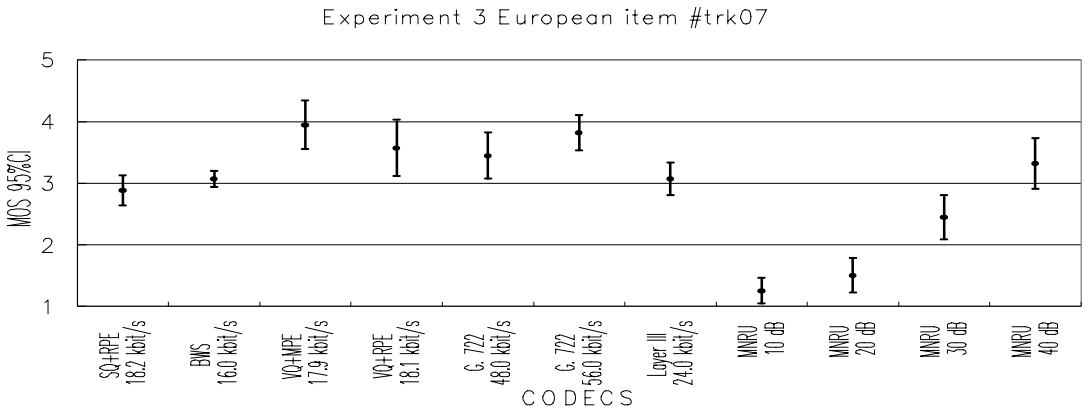
Item 04, Male (German)



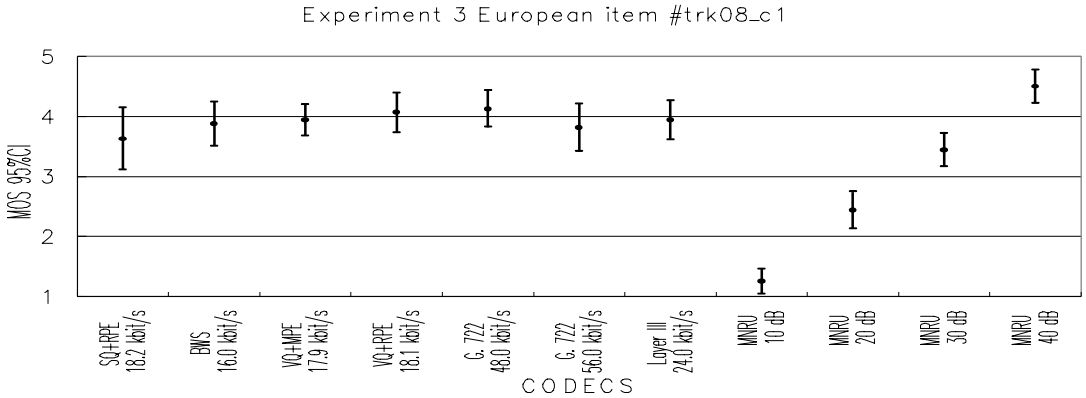
Item 06, Male (English)



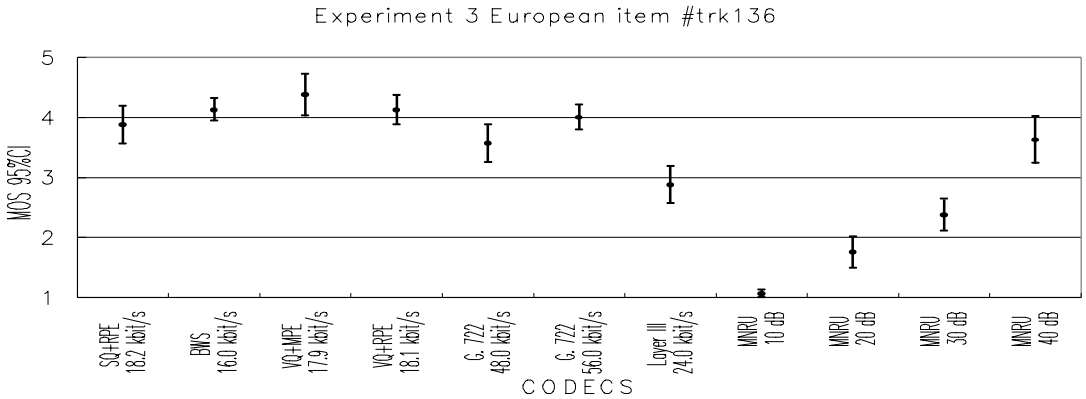
Item 07, Male (English)



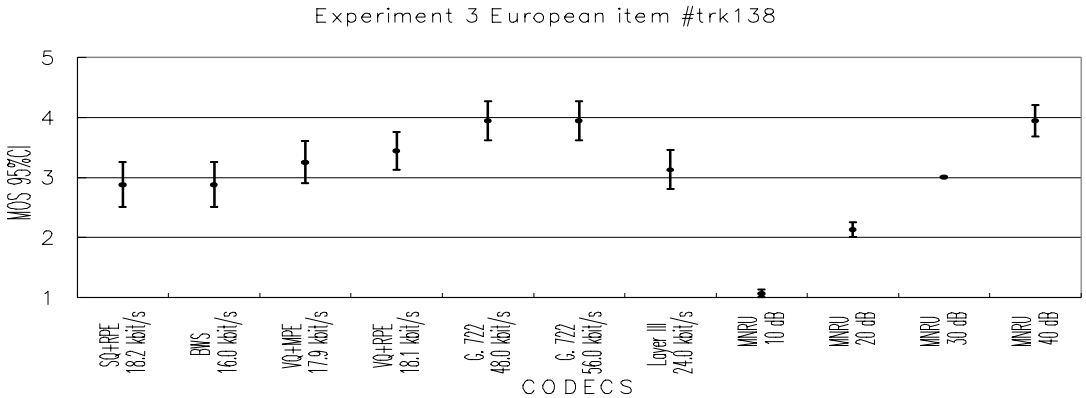
Item 08_c1, Male with car background noise (English)



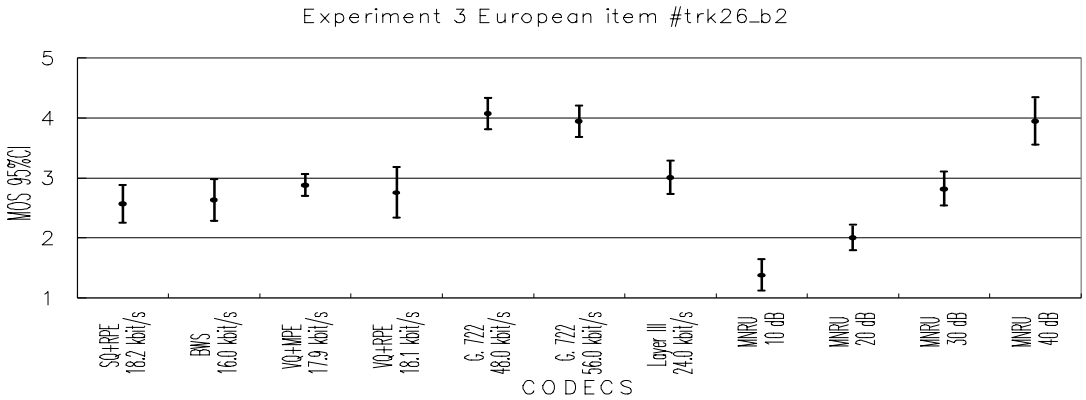
Item 136, Male (Swedish)



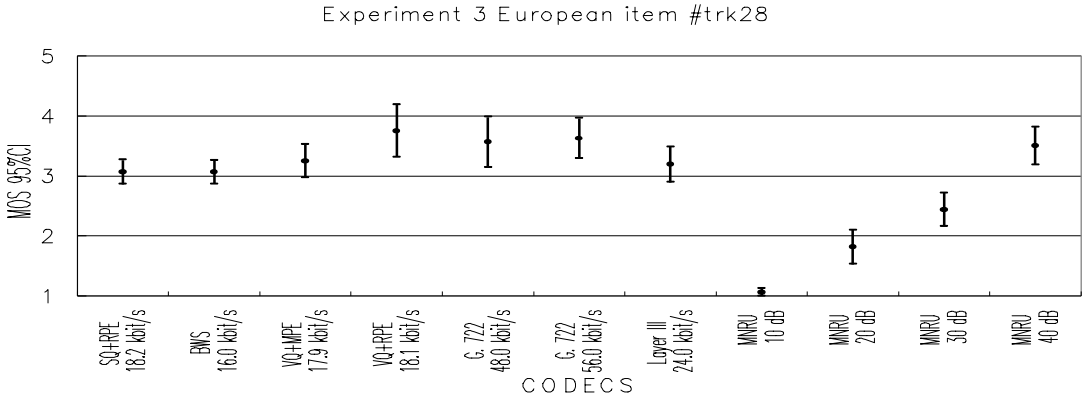
Item 138, Female (Swedish)



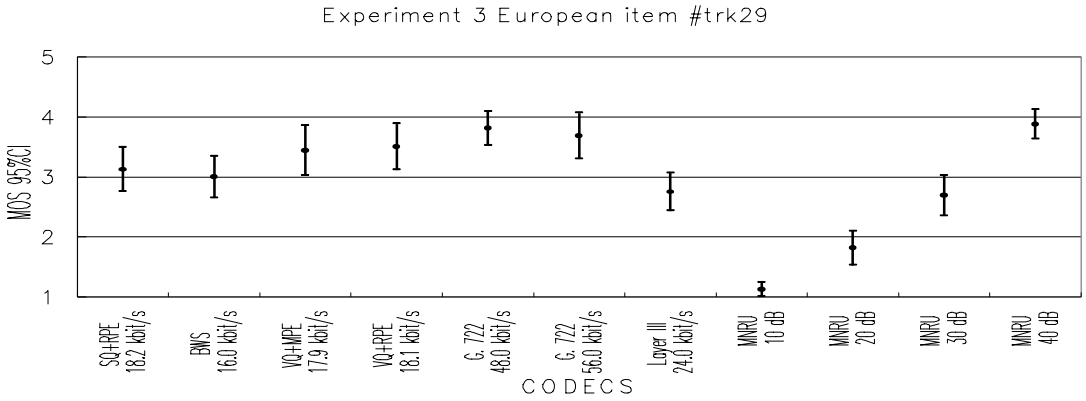
Item 26_b2, Female with babble background noise (German)



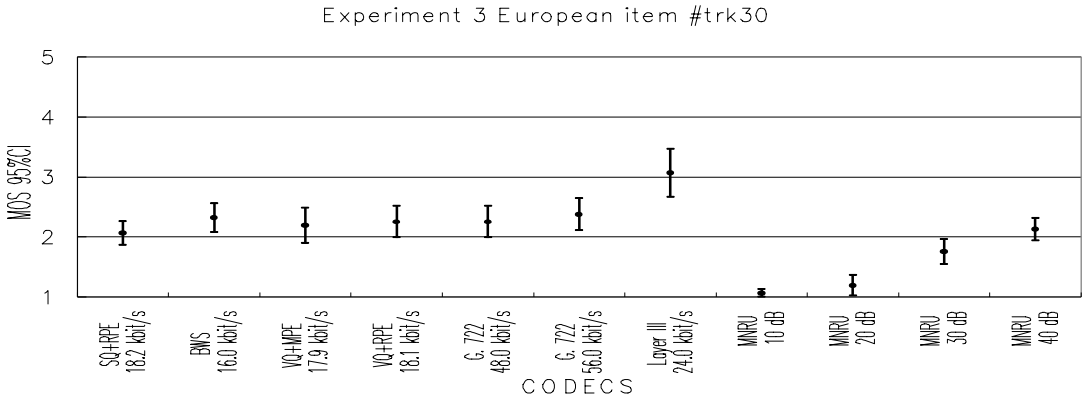
Item 28, Female (German)



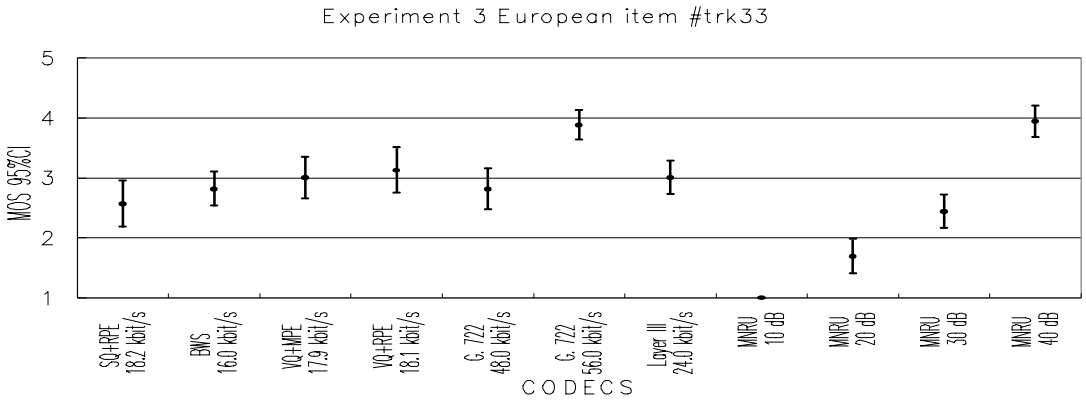
Item 29, Female (German)



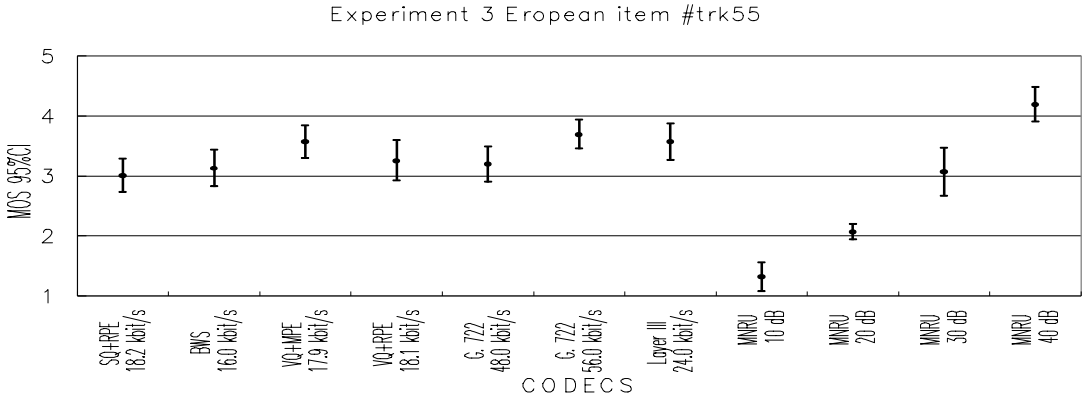
Item 30, Female (English)



Item 33, Female (English)



Item 55, Background music (English)



Item 83, Classical music

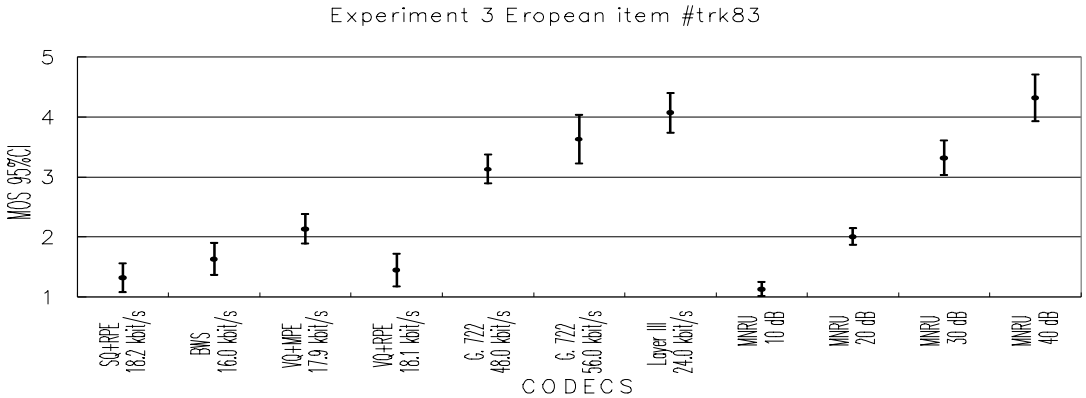
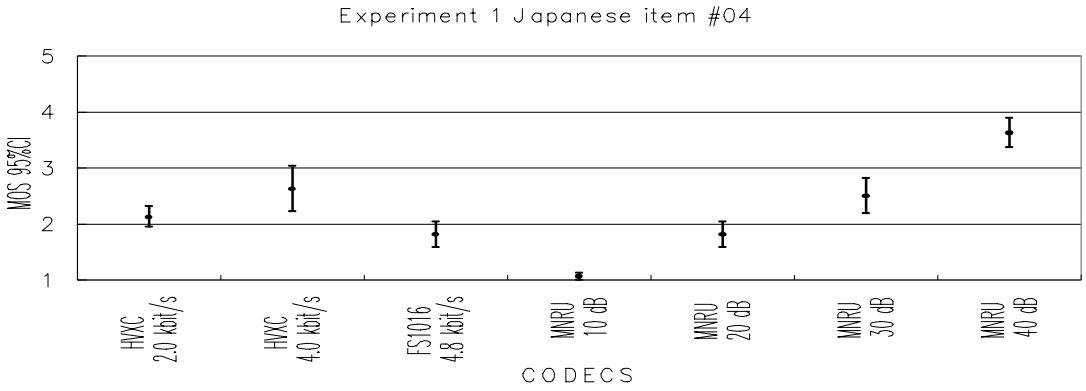


Figure 18. Item by item results of the listening test 3 (WB-CELP).

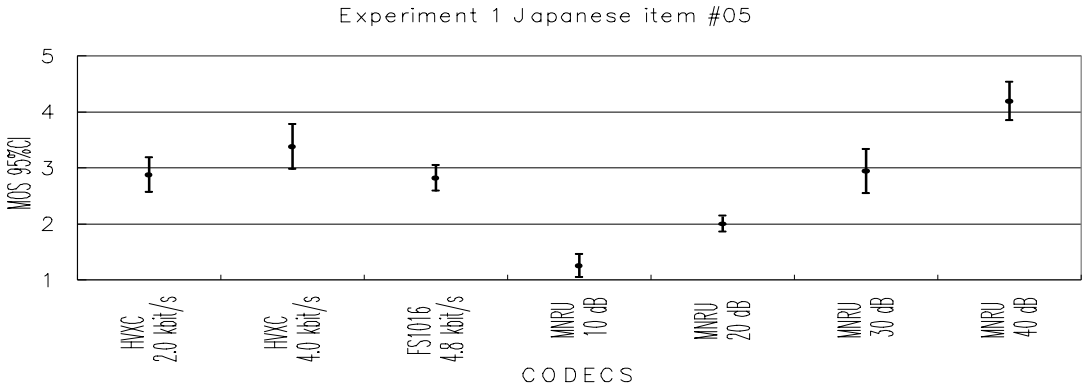
9.3.3 NTT site

The performance of each coder in experiments 1, 2 and 3 are shown graphically in Figures 19, 20 and 21, respectively.

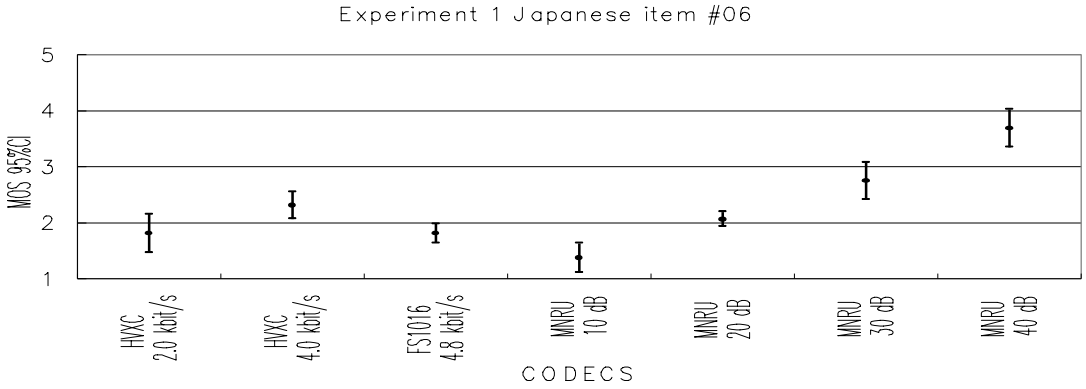
Item 04, Female (Japanese)



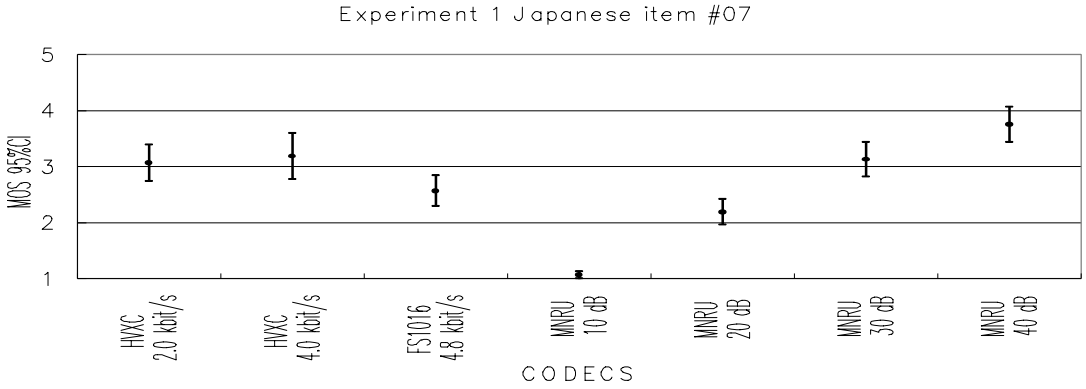
Item 05, Male (Japanese)



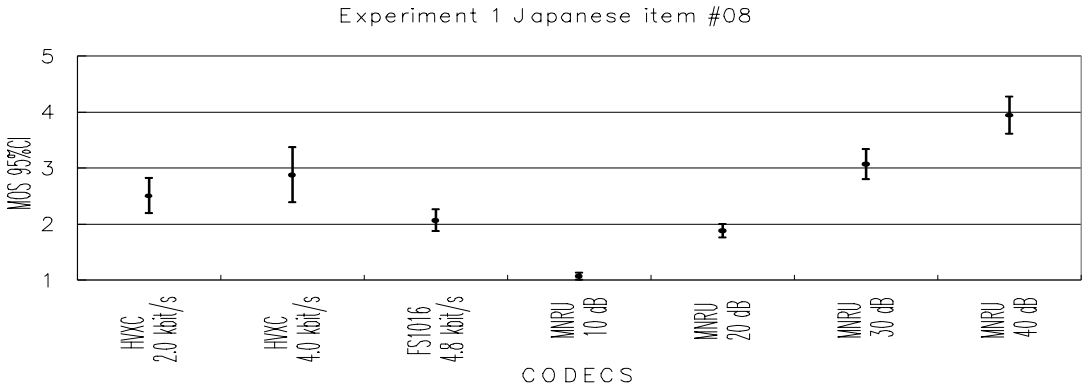
Item 06, Female (Japanese)



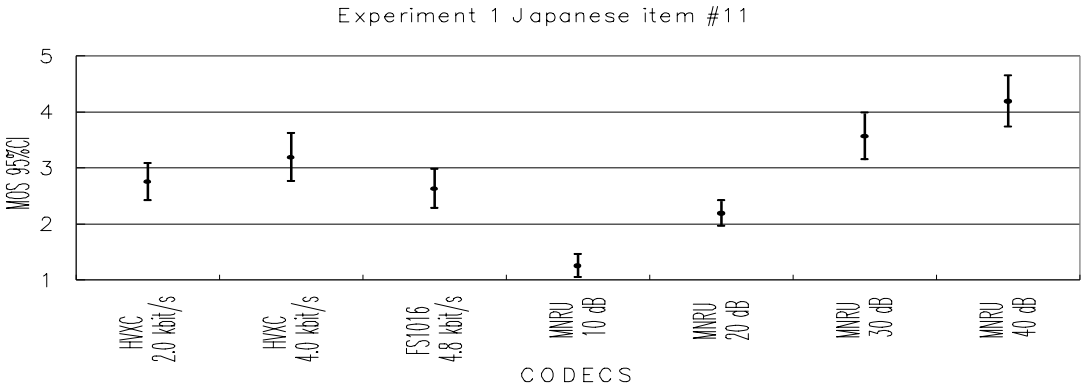
Item 07, Male (Japanese)



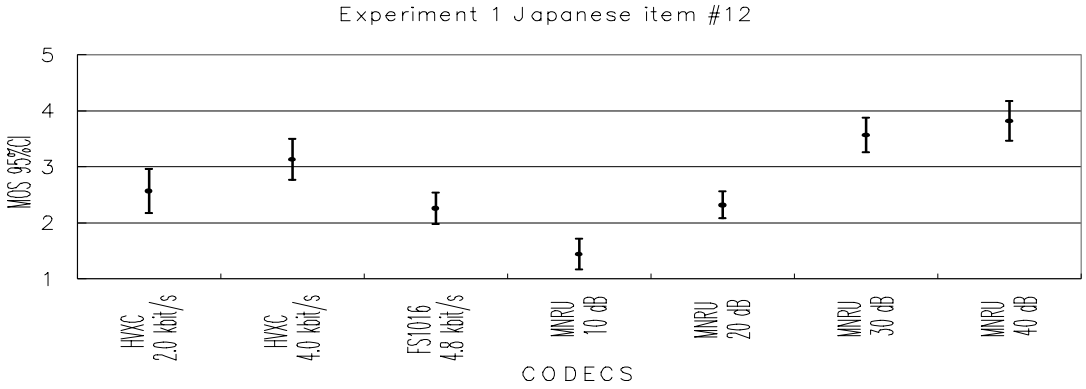
Item 08, Female (Japanese)



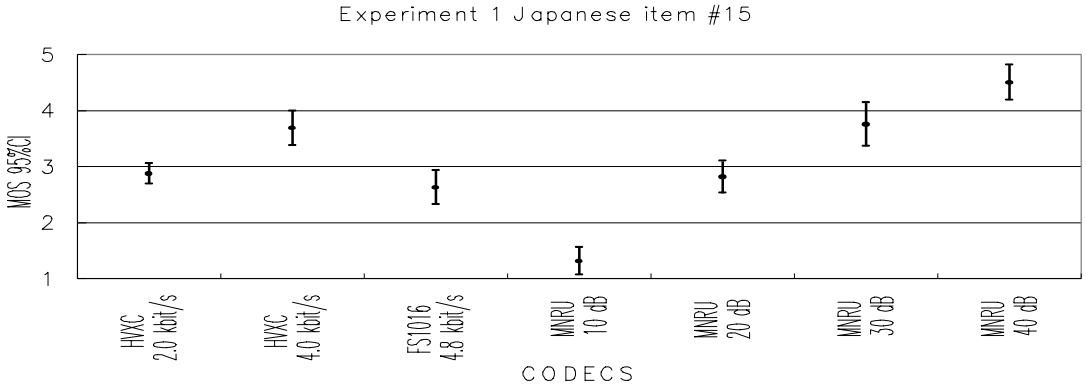
Item 11, Male (Japanese)



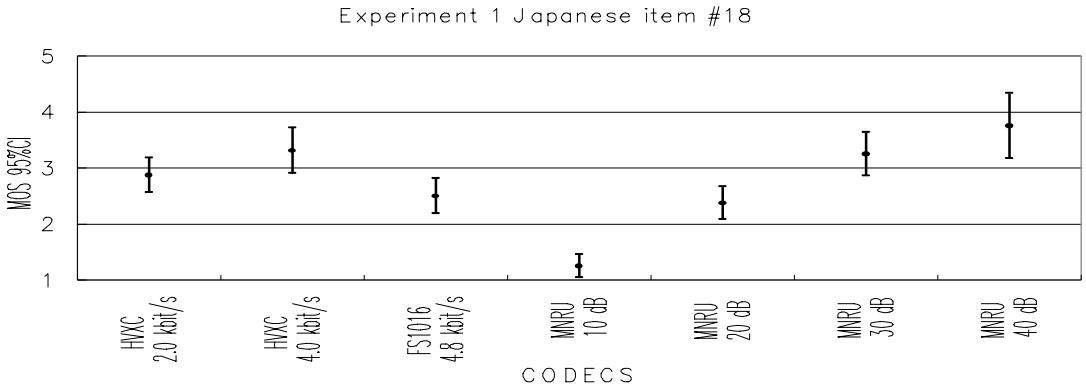
Item 12, Female (Japanese)



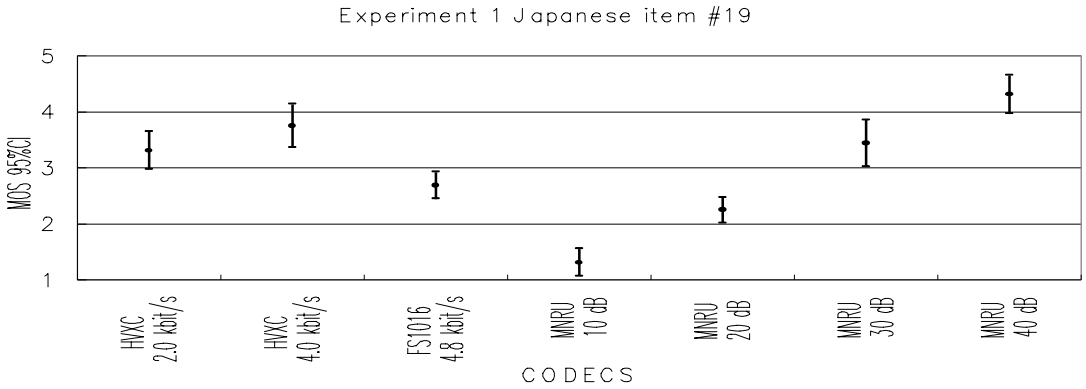
Item 15, Male (Japanese)



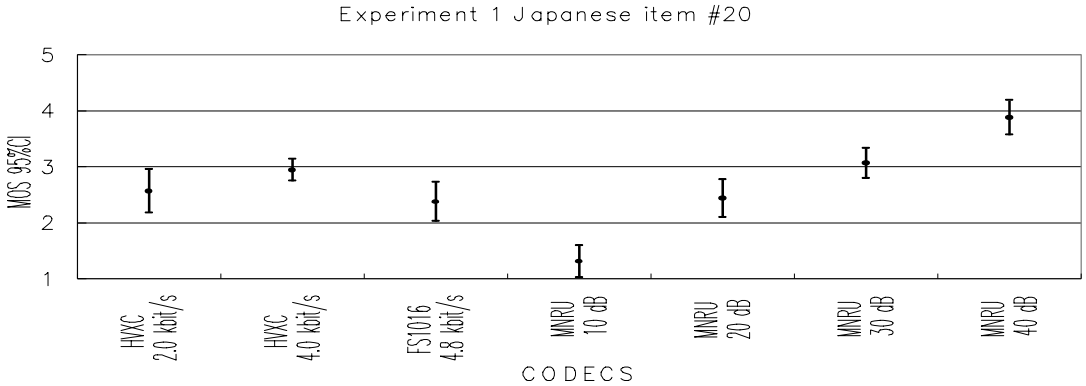
Item 18, Female (Japanese)



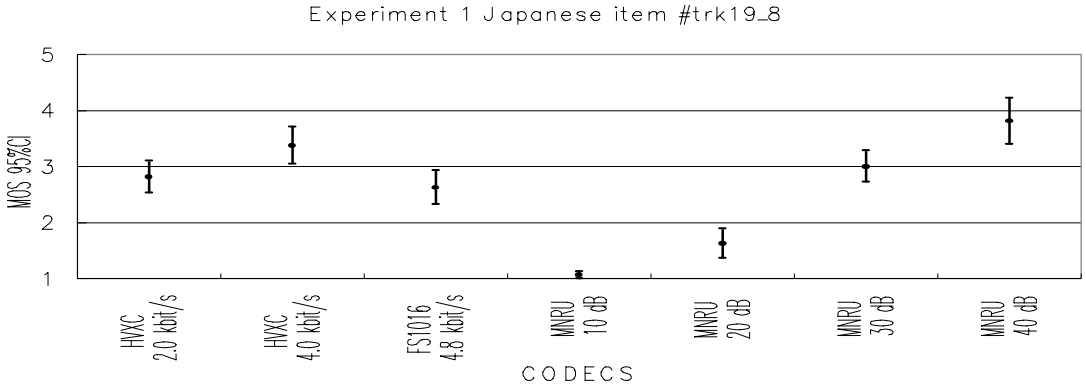
Item 19, Male (Japanese)



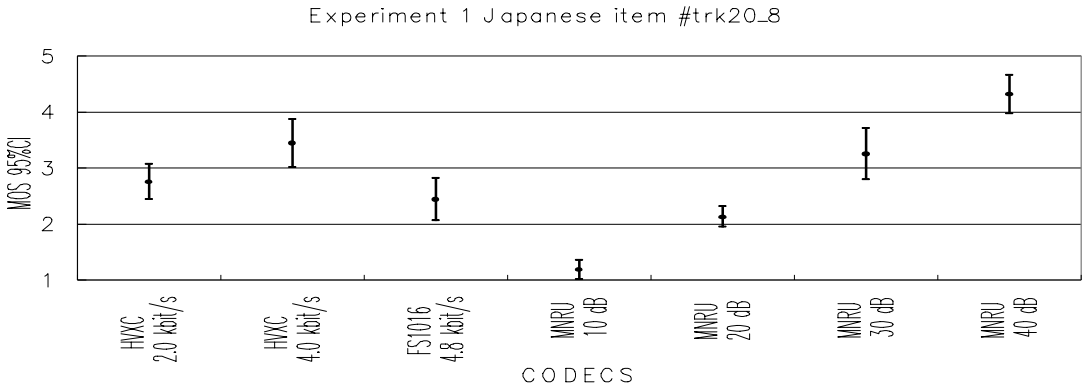
Item 20, Female (Japanese)



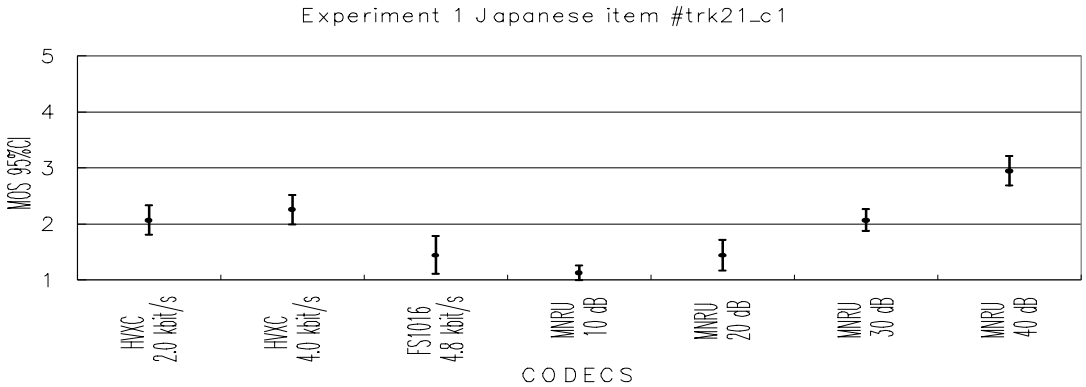
Item trk19_8, Male (Japanese)



Item trk20_8, Male (Japanese)



Item trk21_c1, Male with car background noise (Japanese)



Item 20_8, Male (Japanese)

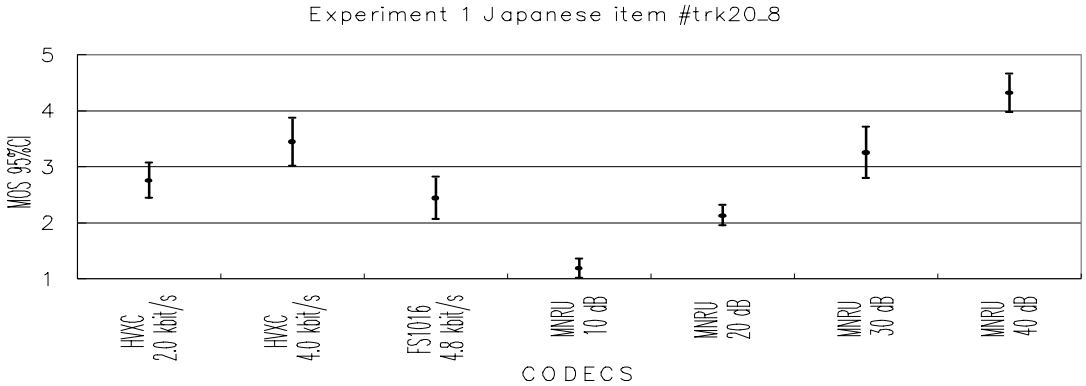
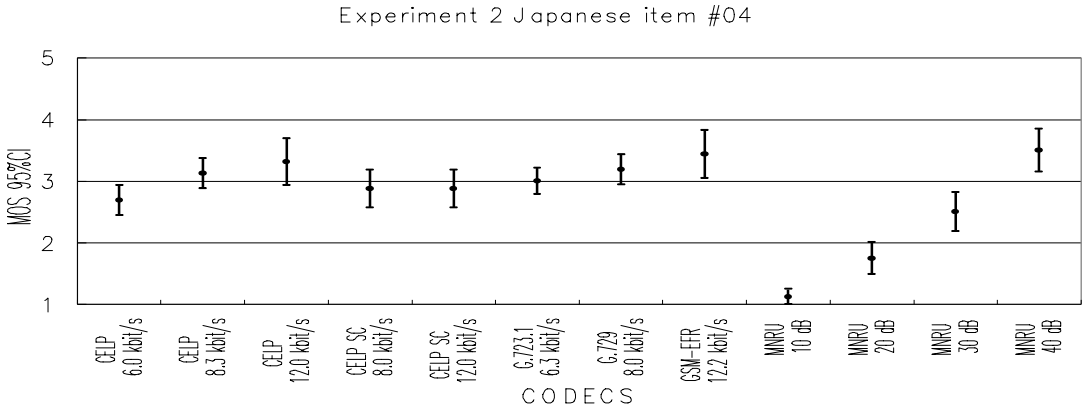
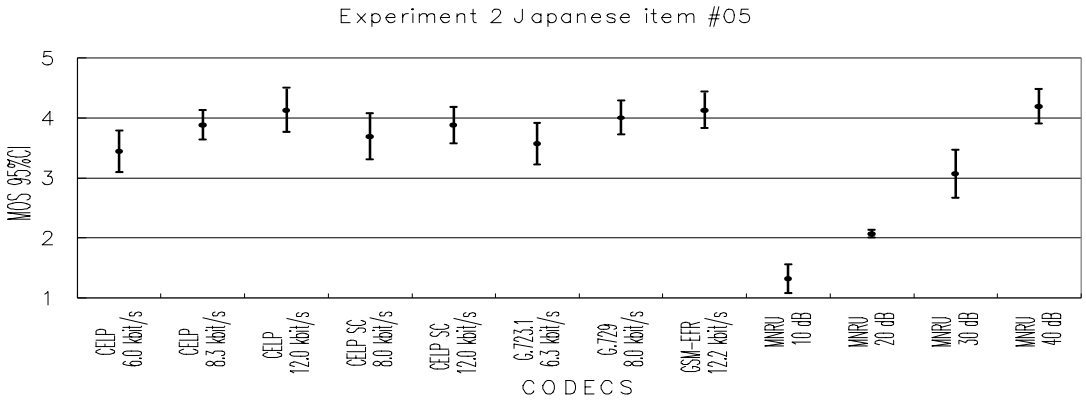


Figure 19. Item by item results of the listening test 1 (PARAMETRIC).

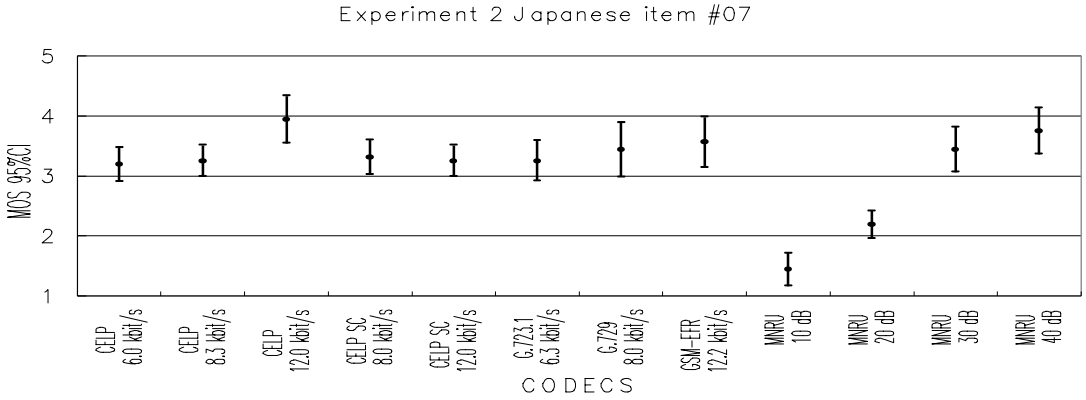
Item 04, Female (Japanese)



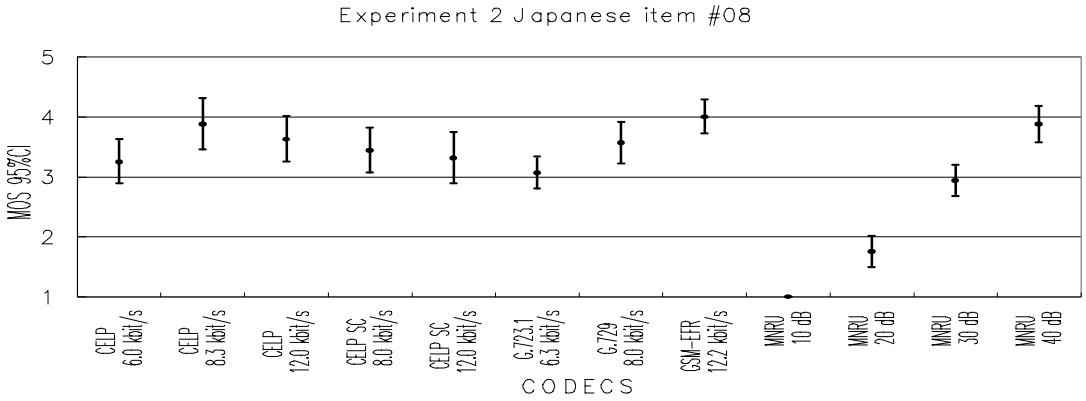
Item 05, Male (Japanese)



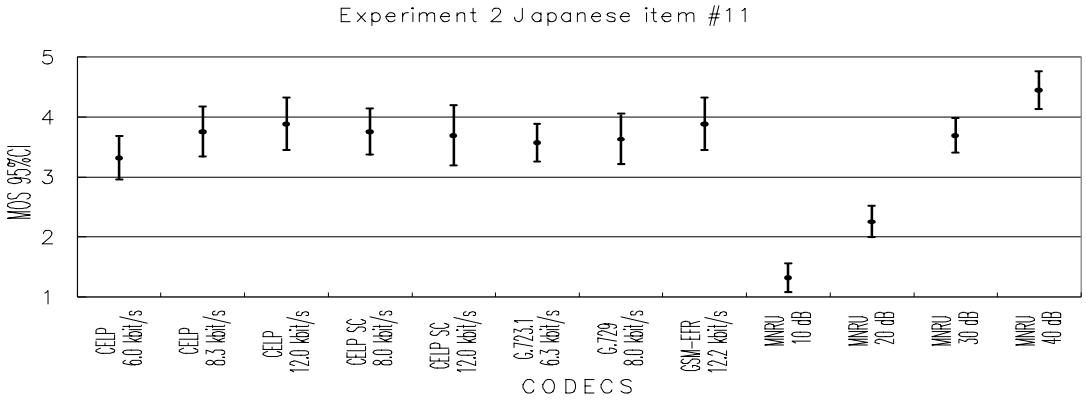
Item 07, Male (Japanese)



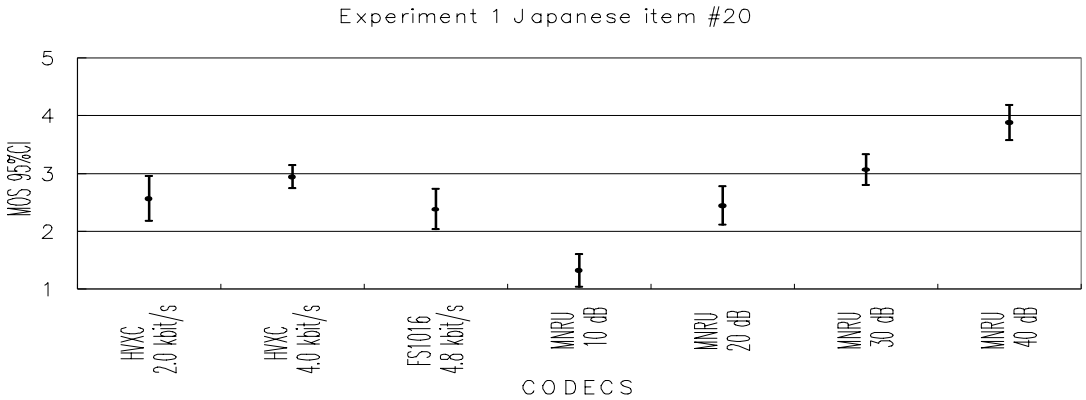
Item 08, Female (Japanese)



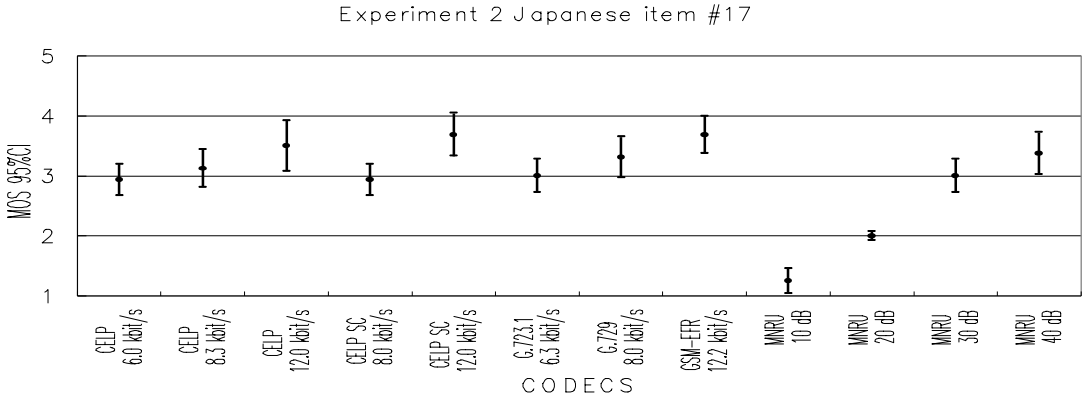
Item 11, Male (Japanese)



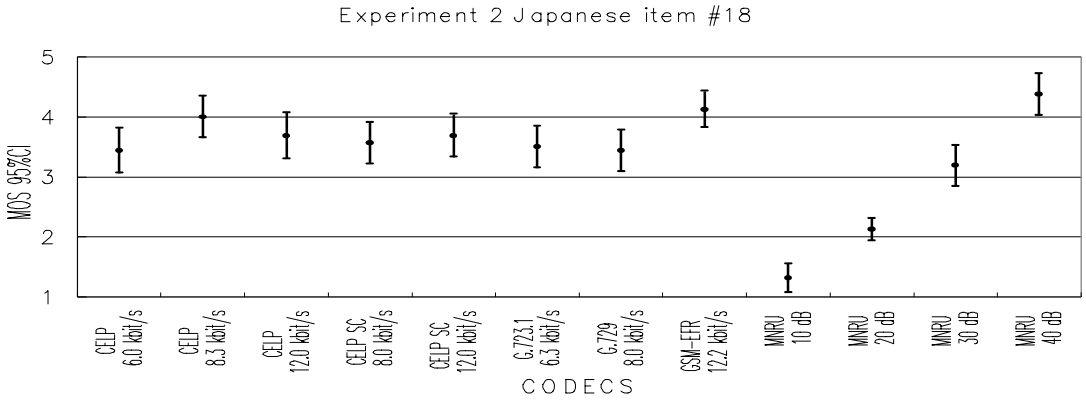
Item 20, Female (Japanese)



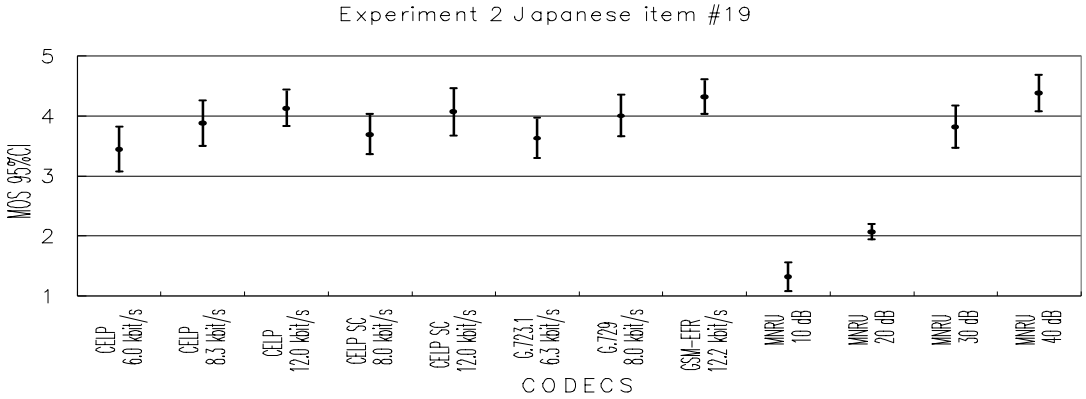
Item 17, Male (Japanese)



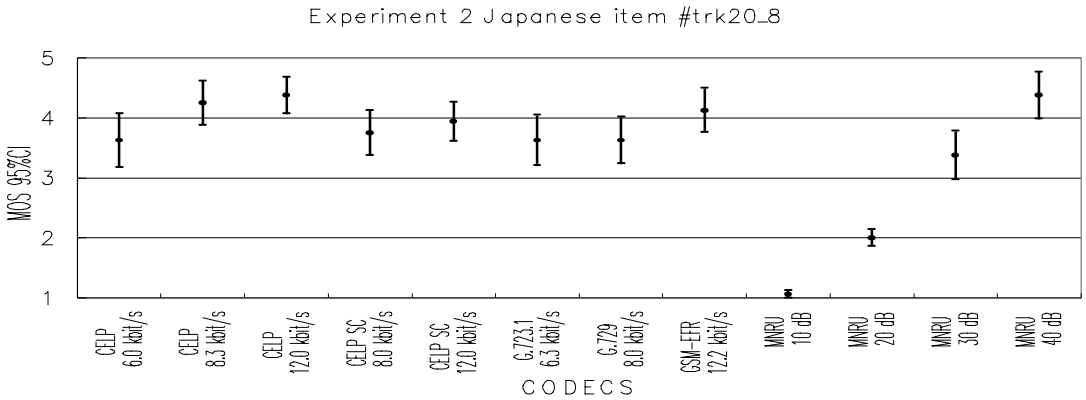
Item 18, Female (Japanese)



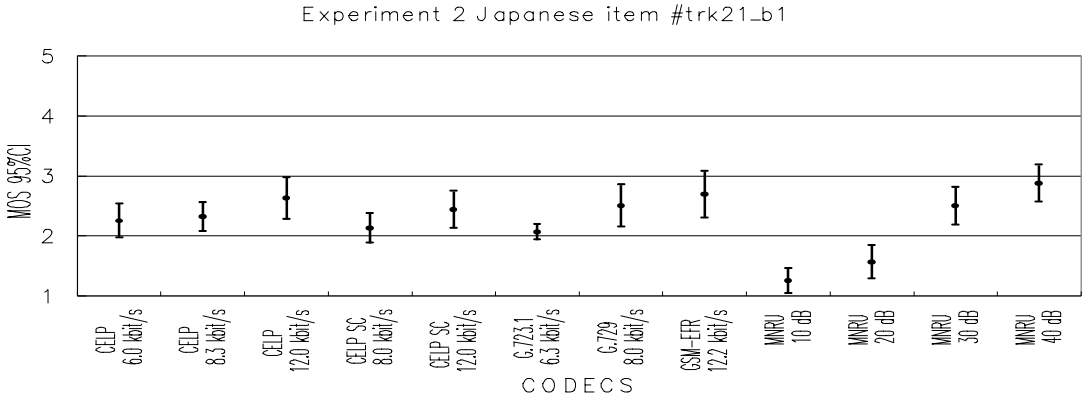
Item 19, Male (Japanese)



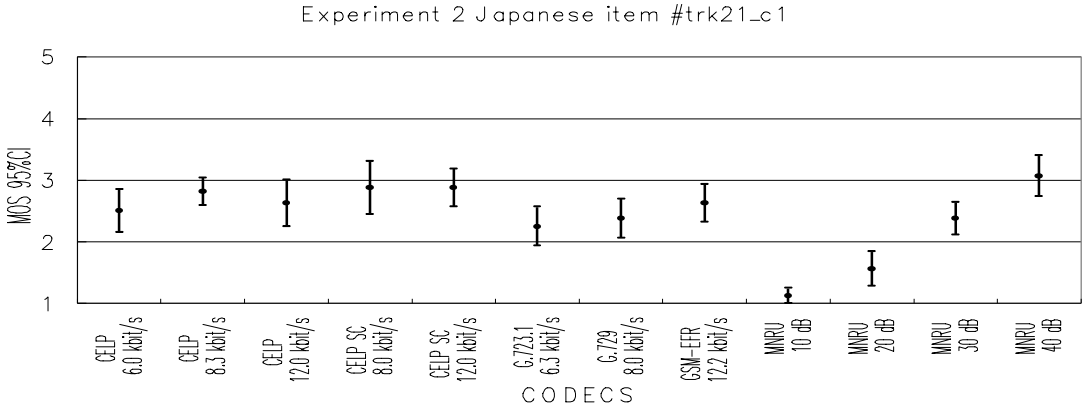
Item trk20_8, Male (Japanese)



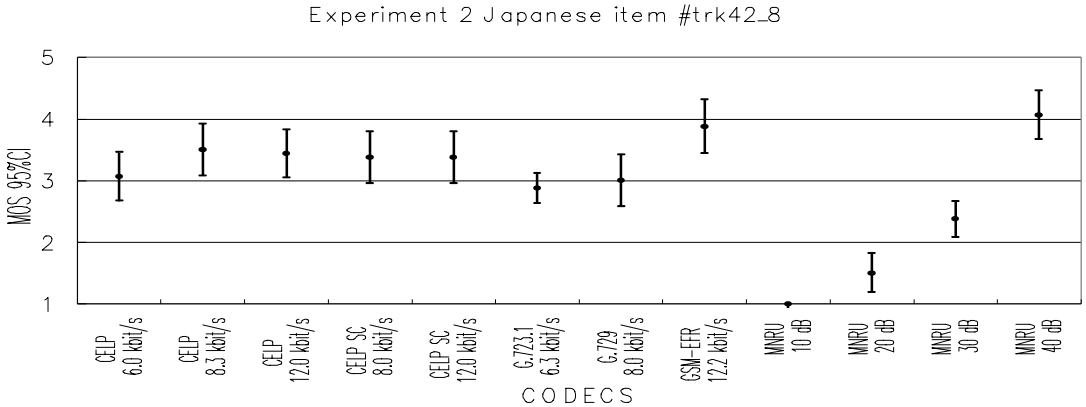
Item trk21_b1, Male with babble background noise (Japanese)



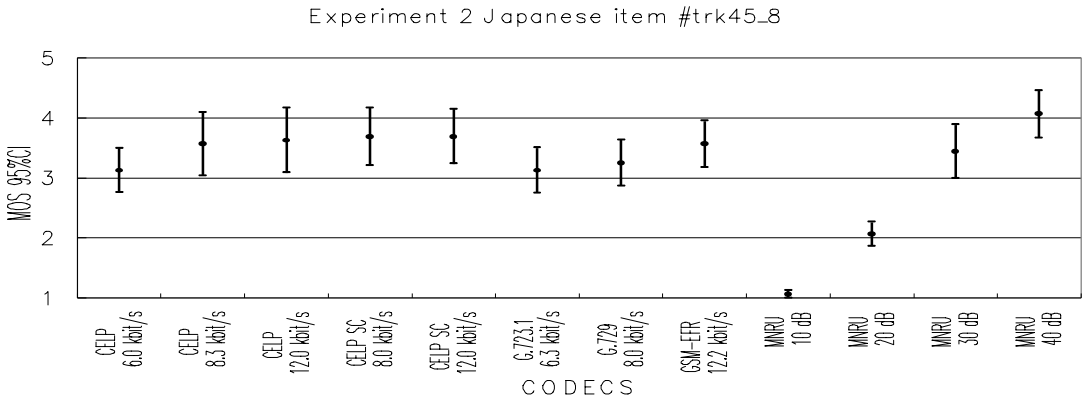
Item trk21_c1, Male with car background noise (Japanese)



Item trk42_8, Female (Japanese)



Item trk45_8, Female (Japanese)



Item trk69_8, Female with background music (Japanese)

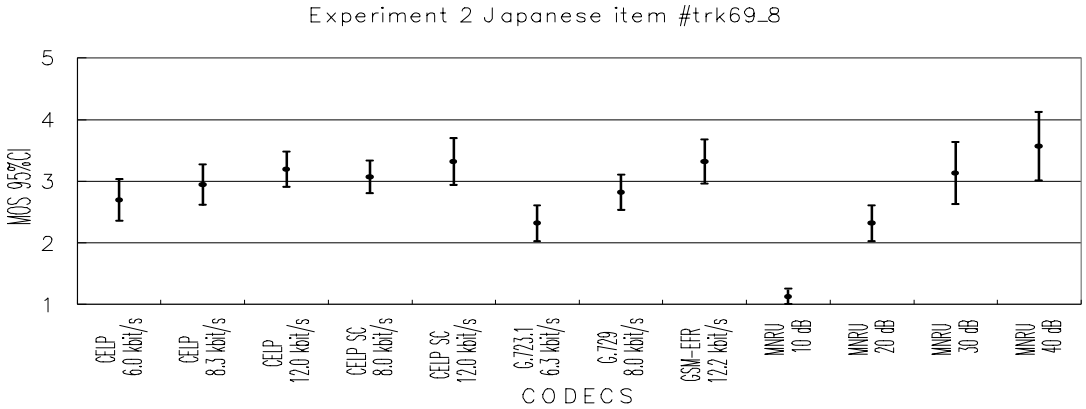
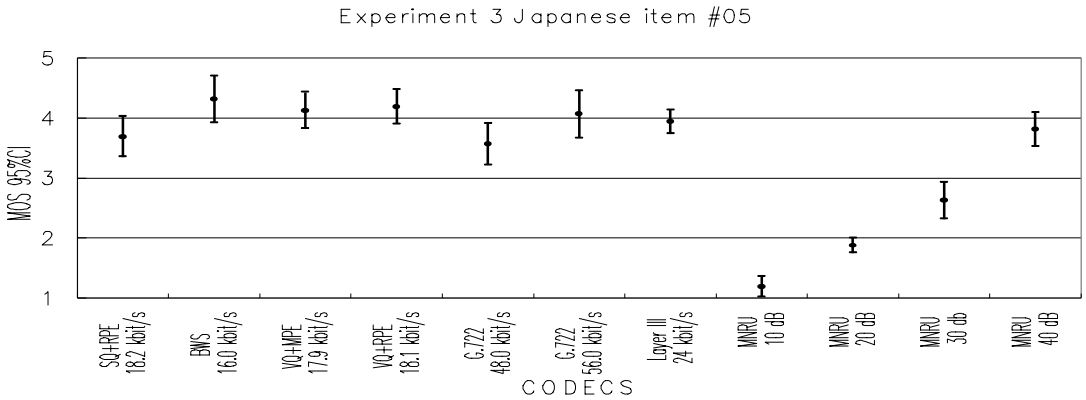
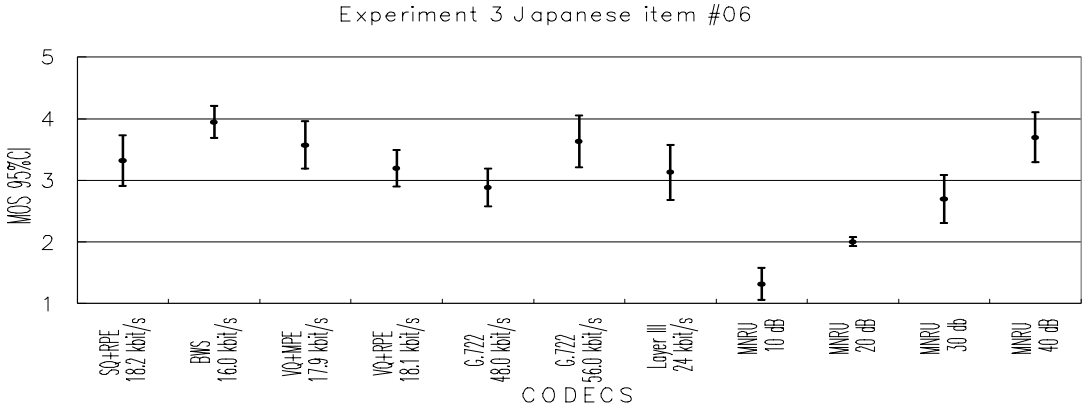


Figure 20. Item by item results of the listening test 2 (NB-CELP).

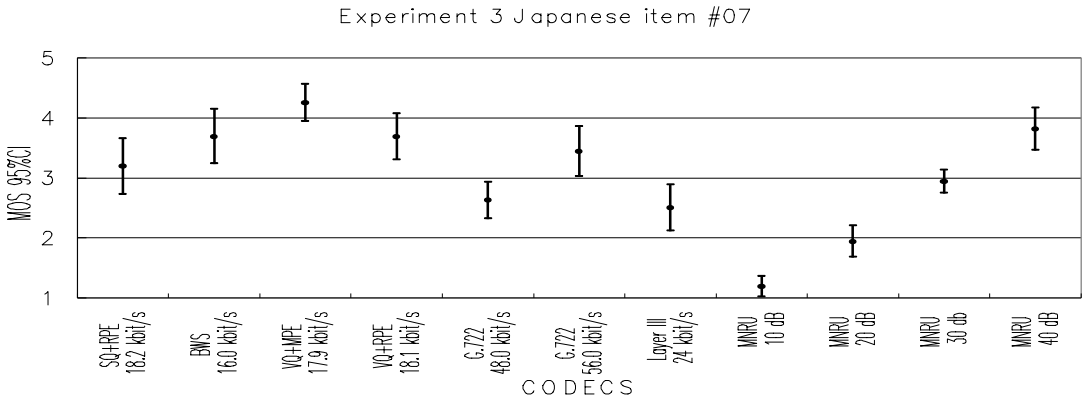
Item 05, Male (Japanese)



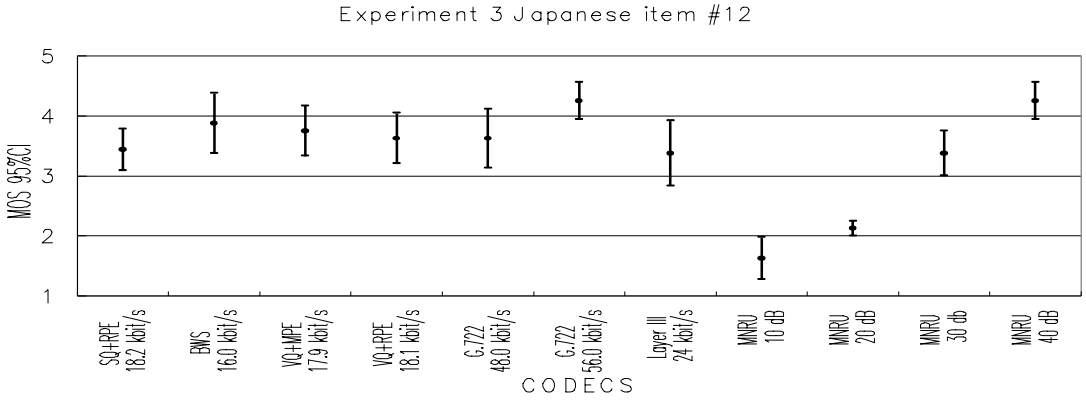
Item 06, Female (Japanese)



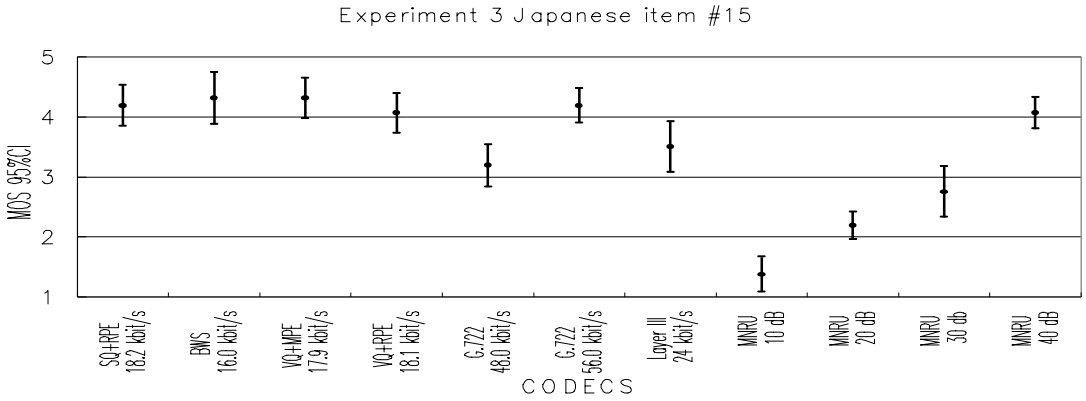
Item 07, Male (Japanese)



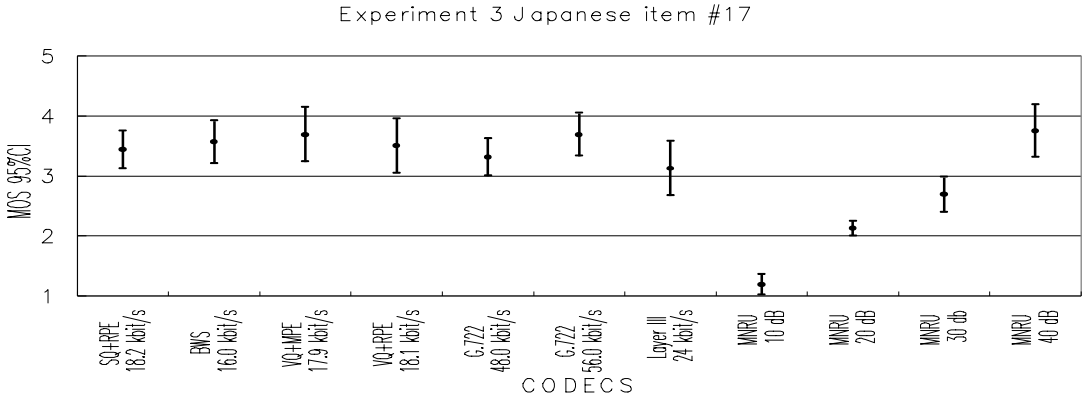
Item 12, Female (Japanese)



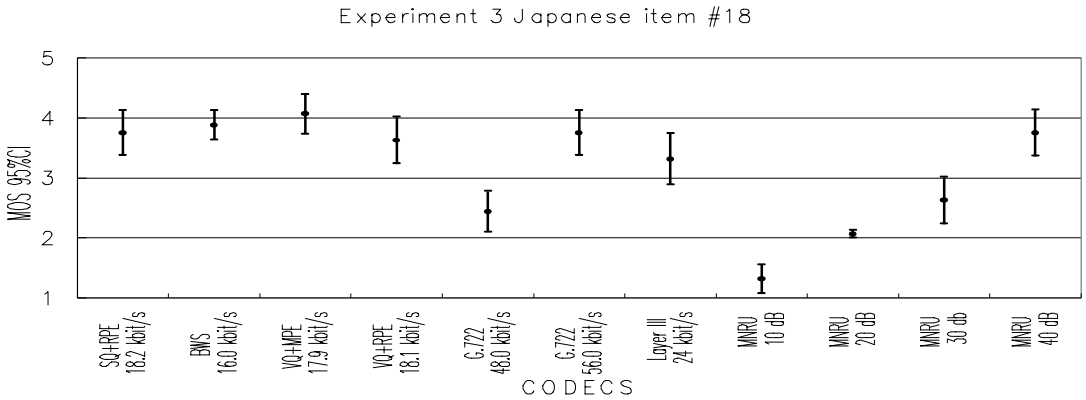
Item 15, Male (Japanese)



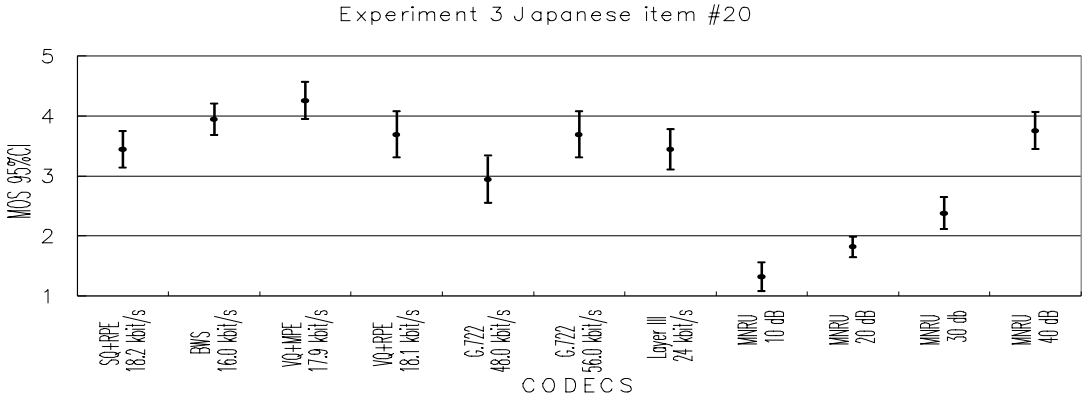
Item 17, Male (Japanese)



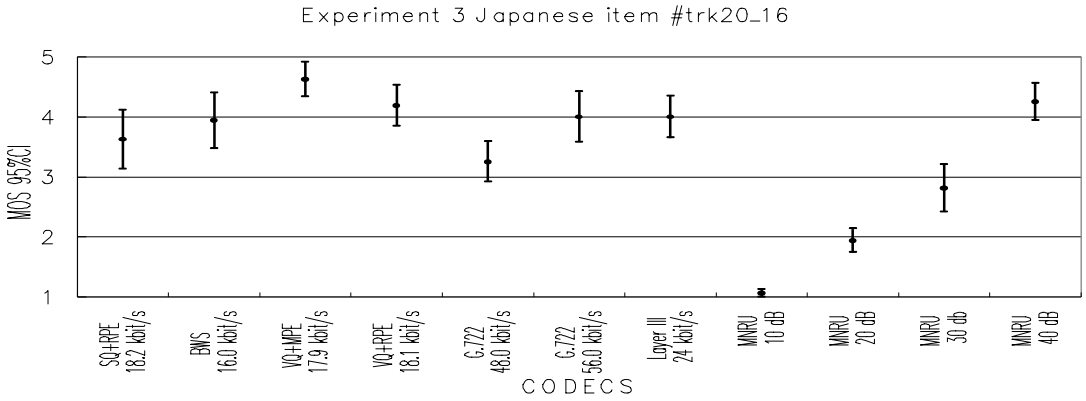
Item 18, Female (Japanese)



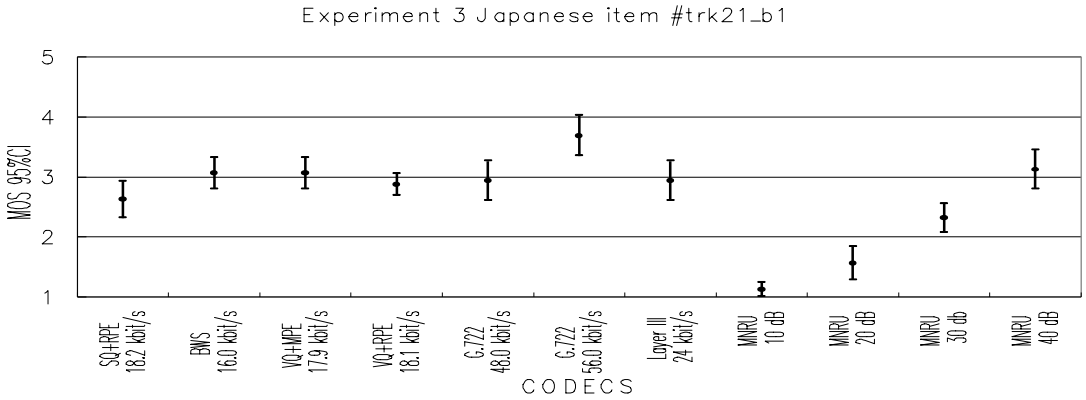
Item 20, Female (Japanese)



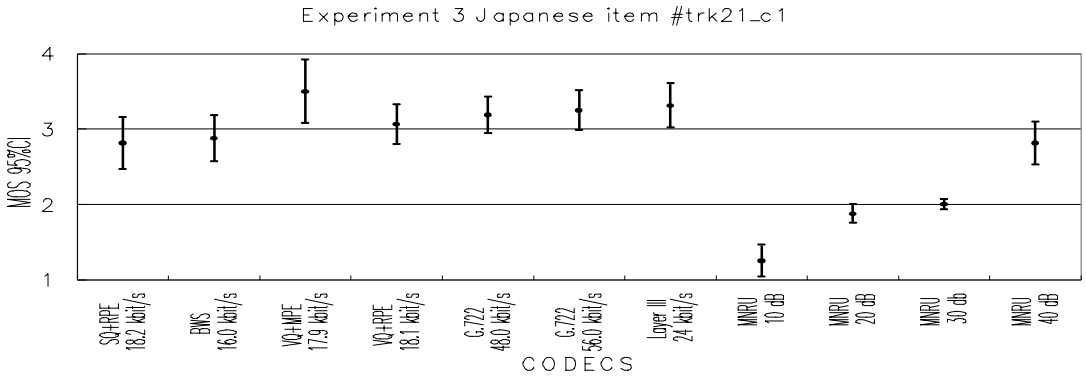
Item trk20_16, Male (Japanese)



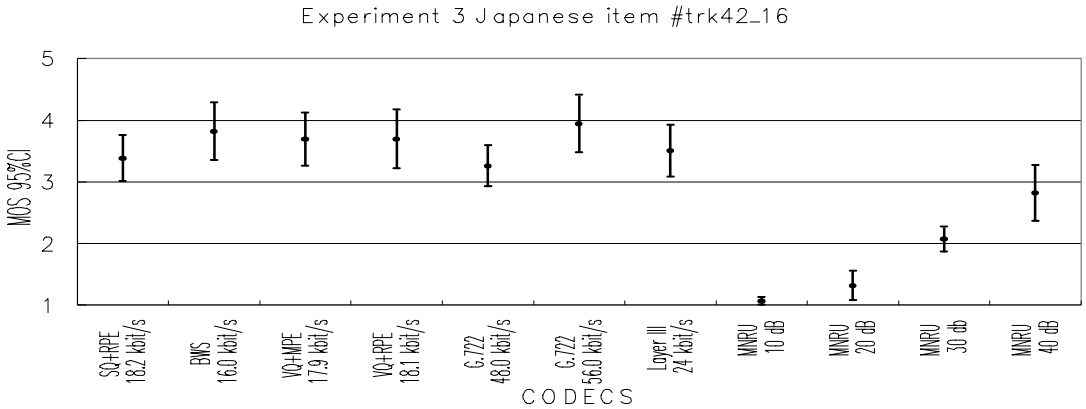
Item trk21_b1, Male with babble background noise (Japanese)



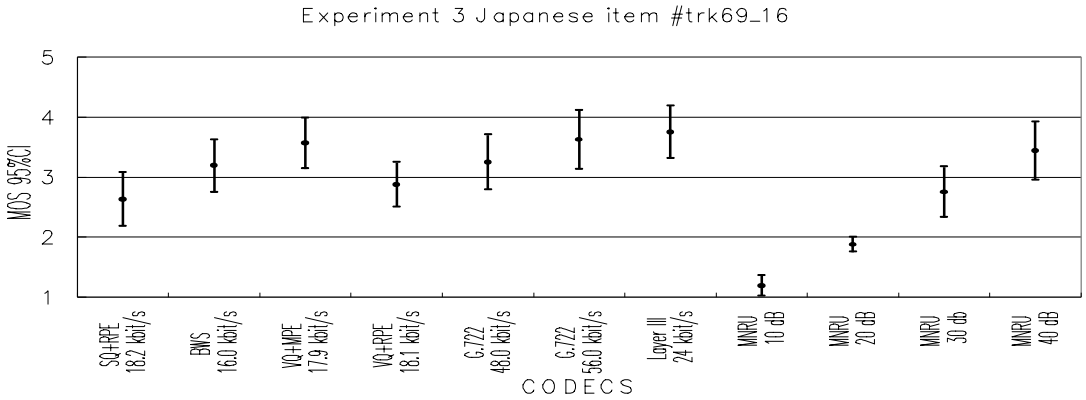
Item trk21_c1, Male with car background noise (Japanese)



Item 42_16, Female (Japanese)



Item 69_16, Female with background music (Japanese)



Item 83, Classical music

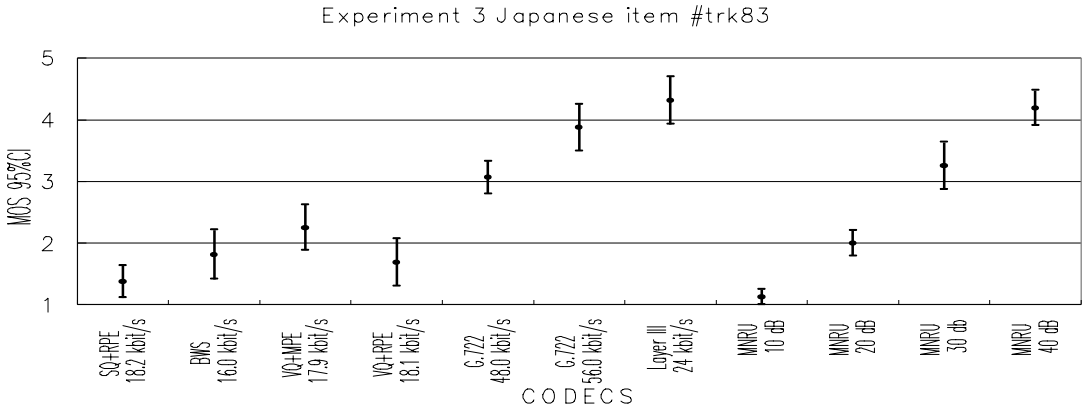


Figure 21. Item by item results of the listening test 3 (WB-CELP).

10 References

[1] ISO/IEC JTC1/SC29/WG11/N2277 MPEG-4 Audio verification tests specifications - speech part, July 1998.

11 Appendix A

This appendix describes the test excerpts used in the listening test. The material was originally collected for the NADIB tests.

Japanese items:

	Pre-selected	Parametric	NB-CELP	WB-CELP
Clean Male	05	X	X	X
	07	X	X	X
	11	X	X	-
	15	X	-	X
	17	-	X	X
	19	X	X	-
	trk19_8	X	-	-
	trk20_8	X	X	X
Clean Female	04	X	X	-
	06	X	-	X
	08	X	X	-
	12	X	X	X
	18	X	X	X
	20	X	-	X
	trk42_8	X	X	X
	trk45_8	-	X	-
Bubble Noise	trk21_b1 (male)	-	X	X
	trk43_b3 (female)	-	-	-
Car Noise	trk21_c1 (male)	X	X	X
	trk43_c2 (female)	-	-	-
Back Music	trk66_8 (female)	-	-	-
	trk67_8 (male)	-	-	-
	trk68_8 (female)	-	-	-
	trk69_8 (female)	-	X	X
Music	trk82 (Classic)	-	-	-
	trk83 (Classic)	-	-	X
	trk114 (English Song)	-	-	-
	trk140 (Swedish Pop)	-	-	-

European items:

	Pre-selected	Parametric	NB-CELP	WB-CELP
Clean Male	trk02(German)	X	X	X
	trk03(German)	-	-	-
	trk04(German)	X	X	X
	trk05(German)	X	X	-
	trk06(English)	-	X	X
	trk07(English)	X	X	X
	trk08(English)	X	-	-
	trk09(English)	X	-	-
	trk136(Swedish)	X	X	X
Clean Female	trk26(German)	X	X	-
	trk27(German)	X	X	-
	trk28(German)	X	-	X
	trk29(German)	X	X	X
	trk30(English)	-	X	X
	trk31(English)	-	X	-
	trk32(English)	X	-	-
	trk33(English)	X	-	X
	trk138(Swedish)	X	X	X
Babble Noise	trk26_b1 (German female)	-	-	-
	trk26_b2 (German female)	-	X	X
	trk31_b1 (English female)	-	-	-
	trk31_b2 (English female)	-	-	-
Car Noise	trk05_c1 (German male)	-	-	-
	trk08_c1 (English male)	X	X	X
Back Music	trk55 (English)	-	X	X
	trk56 (English)	-	-	-
	trk57 (English)	-	-	-
Music	trk82 (Classic)	-	-	-
	trk83 (Classic)	-	-	X
	trk114 (English Song)	-	-	-
	trk140 (Swedish Pop)	-	-	-