

Tdoc S4-AHQ034

Source:	Qualcomm, Inc.
Title:	Preliminary results of ETSI EG 202-396-3 validation with narrowband dual-microphone noise cancelling terminals.
Document for:	Discussion
Agenda Item:	Unique
Date:	October, 25 th 2011

Summary

- A P.835 experiment was conducted according to the conditions established in Tdoc S4-110756.
- The experiment attempted to replicate the same conditions used for the training and validation of the ETSI method as described in ETSI EG 202.396-2.
- 12 “reference” conditions (taken from the ETSI EG 202.396-2 database) and 20 “test” conditions were used.

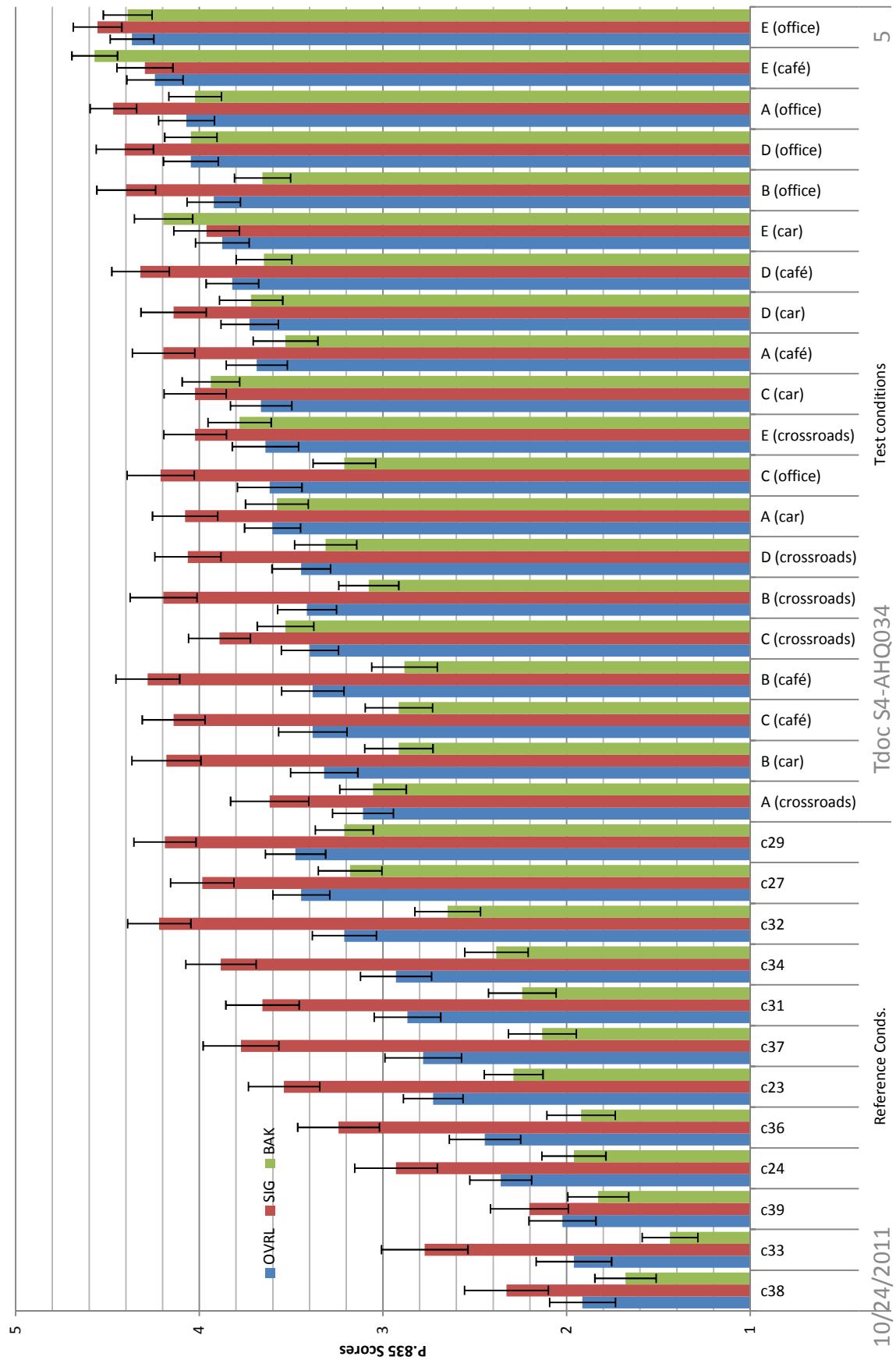
Experimental design

- Used same speech material as ETSI.
- 2 male and 2 female talkers.
- 1 sentence pair per talker.
- Material presented in sentence pairs following ETSI procedure.
- Diotic presentation, 73dB SPL per ear, diffuse field equalization with Sennheiser HD25 headsets.
- 32 Naïve, native American English speakers participants.
- Randomization and blocking performed following guidelines in ITU-T handbook on subjective test procedures.
- Design of experiment resulted in 1 “panel” of 32 listeners listening to a single randomization. Session was divided in 4 blocks. All listeners listened to all samples.
- 32 listeners * 4 votes per condition = 128 votes per condition.

Test Conditions

- 5 dual microphone noise cancelling terminals.
- 4 noise types used (same ones as used in the ETSI LOT):
 - Crossroads
 - Full Size Car 130km/h
 - Office Callcenter
 - Café (Mensa)

Results summary plot



10/24/2011

Tdoc S4-AHQ034

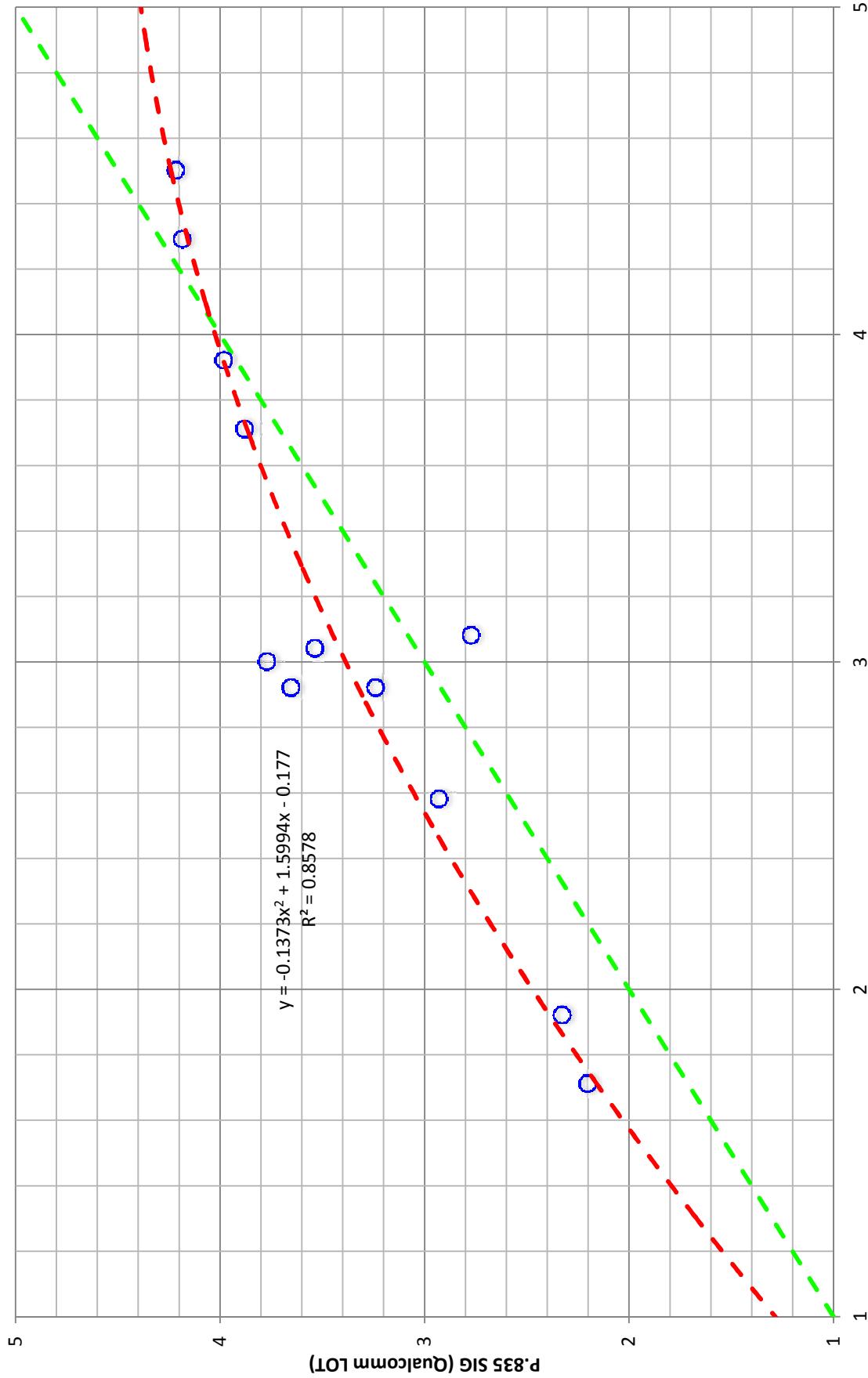
Reference Conds.

Test conditions

Correlation of MOS-LQoS scores between ETSI EG 202.396-2 and Qualcomm experiment

- ETSI EG 202.396-2 did not have a MNRU based reference set. 12 of the conditions used in the ETSI EG 202.396-2 (selected to span the entire MOS scale) were reused in this experiment as “reference” conditions.
- Presentation levels and conditions were supposed to be the same between the two LOTs.
- The correlation between the subjective results of both tests is shown in the following plots.

Qualcomm P.835 SIG x ETSI EG 202.396-2 P.835 SIG (reference set only) (R = 0.926)

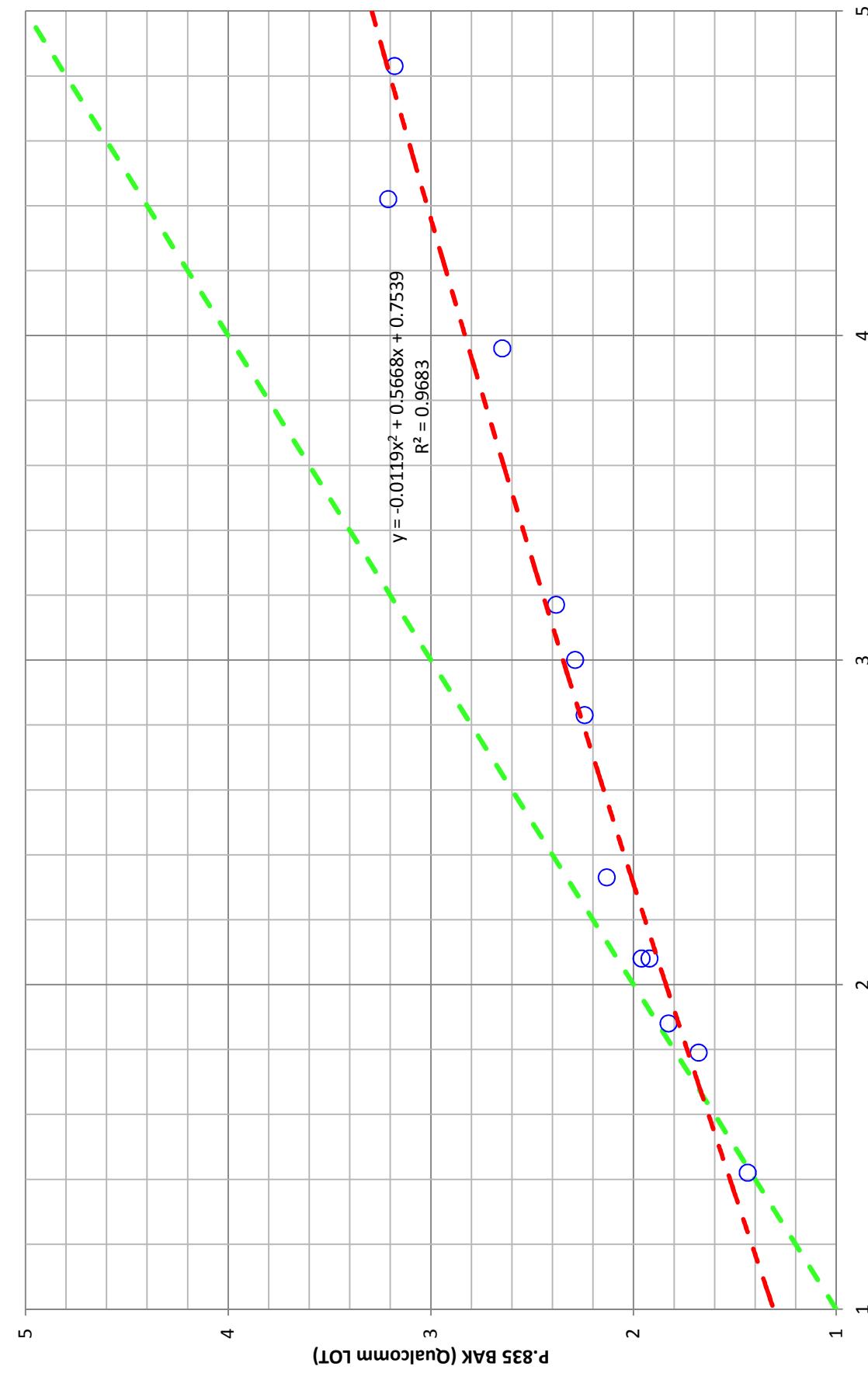


10/24/2011

Tdoc S4-AHQ034

P.835 SIG (ETSI EG 202.396-2 LOT)

Qualcomm P.835 BAK x ETSI EG.202.396-2 P.835 BAK (reference set only)

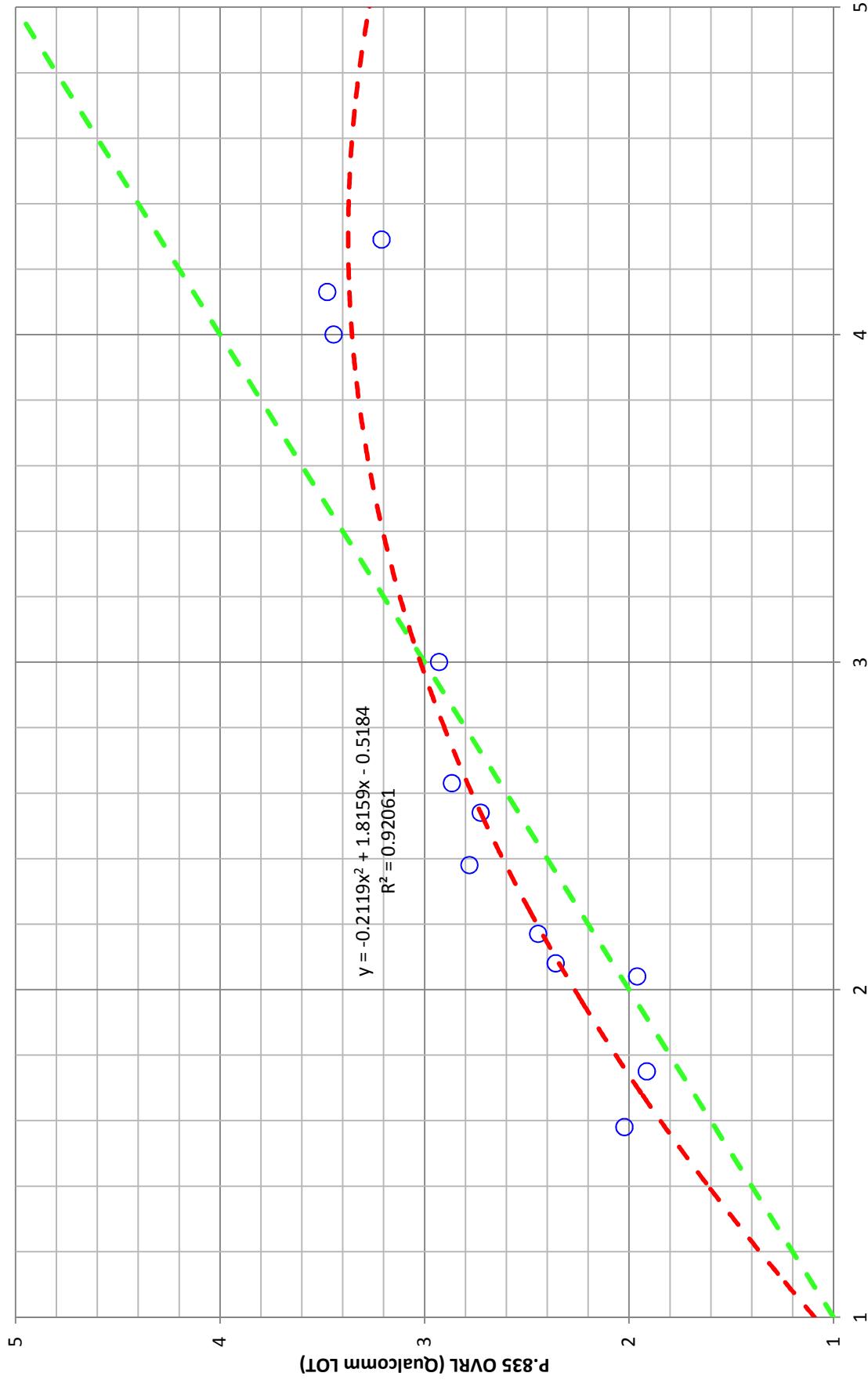


10/24/2011

P.835 BAK (ETSI EG 202.396-2 LOT)
Tdoc S4-AHQ034

8

Qualcomm P.835 BAK x ETSI EG.202.396-2 P.835 BAK (reference set only) (R = 0.959)



10/24/2011

Tdoc S4-AHQ034

P.835 OVR (ETSI EG 202.396-2 LOT)

Preliminary observations

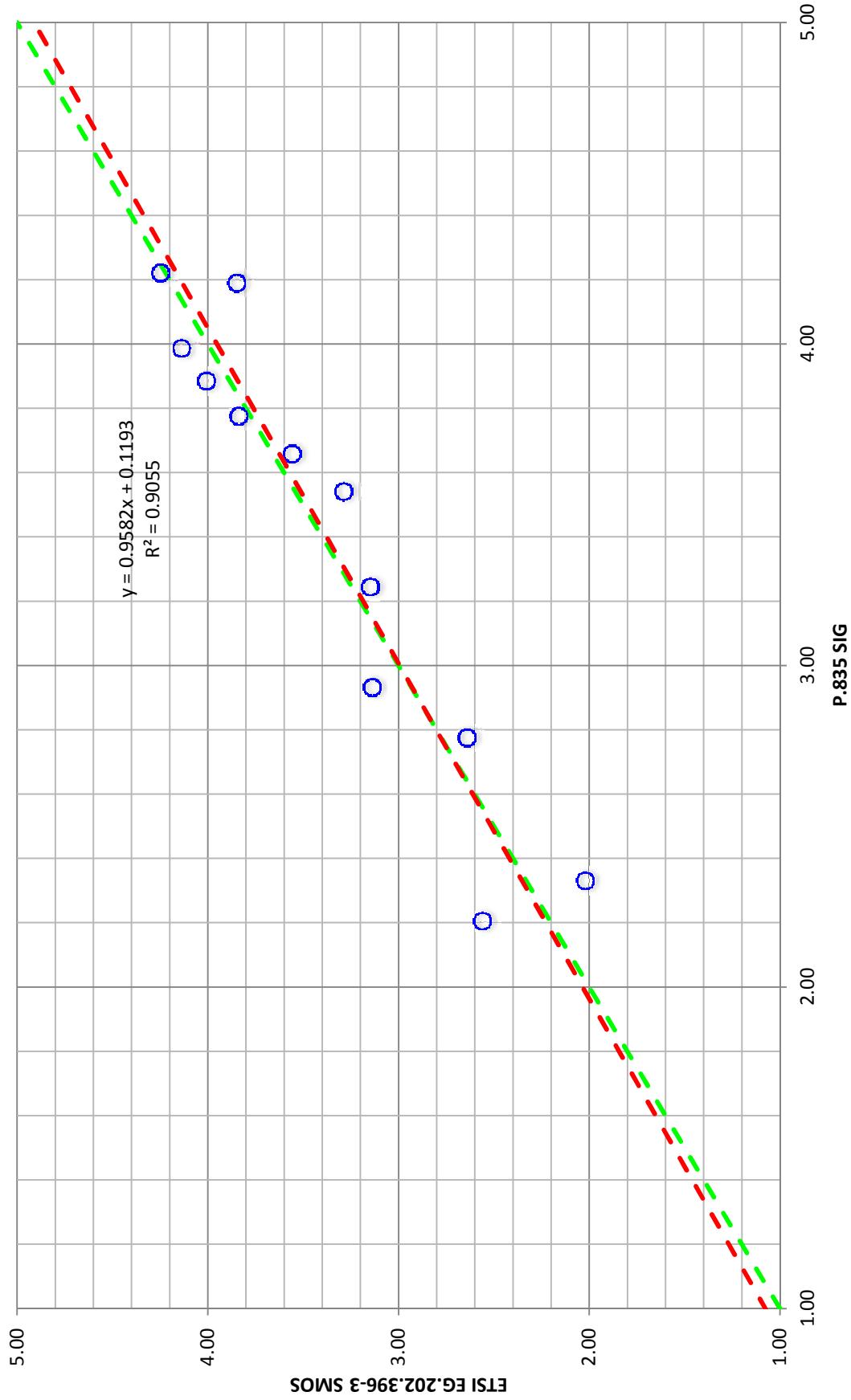
- In the Qualcomm LOT, the “reference” conditions did not succeed in spanning the entire range of the MOS scale as in the original test.
- Correlation between LOTs is generally good but, without mapping, there are significant differences in the BAK scores.
- The differences in BAK scores propagate to the P.835 MOS scores.
- Two hypothesis for the problem:
 - Difference in presentation level calibration procedures lead to significant differences in detectability of noise*
 - The dual microphone noise cancelling terminals introduced a new level of performance previously not experienced in the original tests. The difference in context leads to subjects scoring the reference conditions lower now.
- In general, the absence of a “clean speech” condition in the reference set seems to be an issue with the existing ETSI database as the test is not properly anchored at the top of the scale.

* in P.835 an answer of 5 for BAK is absolute and objective (“*the BACKGROUND in this sample was NOT NOT/CLEAN*”) and the method is sensitive to presentation level differences.

Correlation of Qualcomm experiment and objective scores (ETSI EG.202.396-3) for reference set

- The ETSI EG.202.396-3 method was applied on the concatenated sentences and the resulting objective score is compared with the subjects mean opinion score.
- The correlation for the reference set is presented first.

Qualcomm P.835 SIG x ETSI EG 202.396-3 S-MOS (reference set only) (R=0.951)

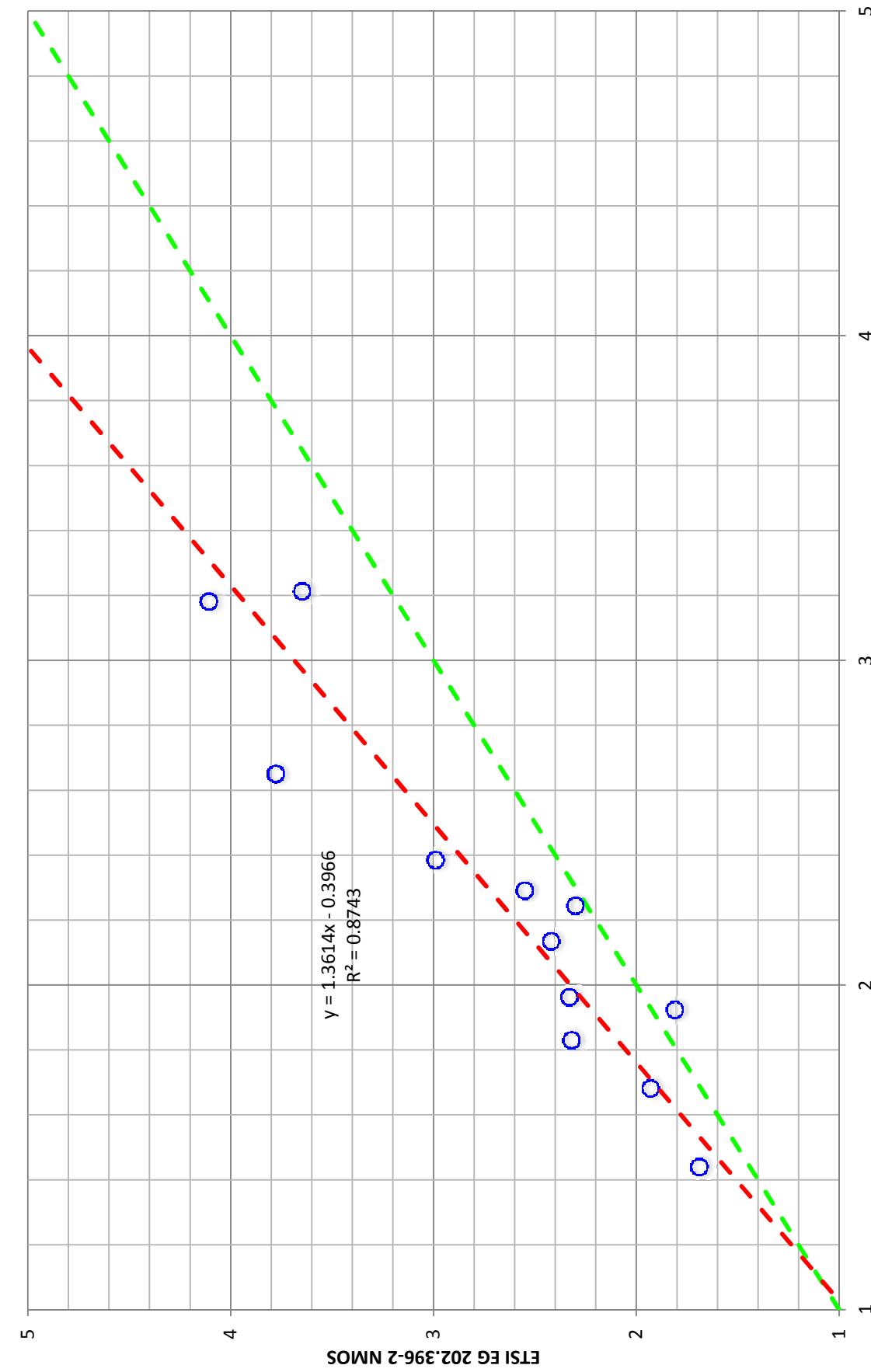


10/24/2011

Tdoc S4-AHQ034

12

Qualcomm P.835 BAK x ETSI EG 202.396-3 N-MOS (reference set only)



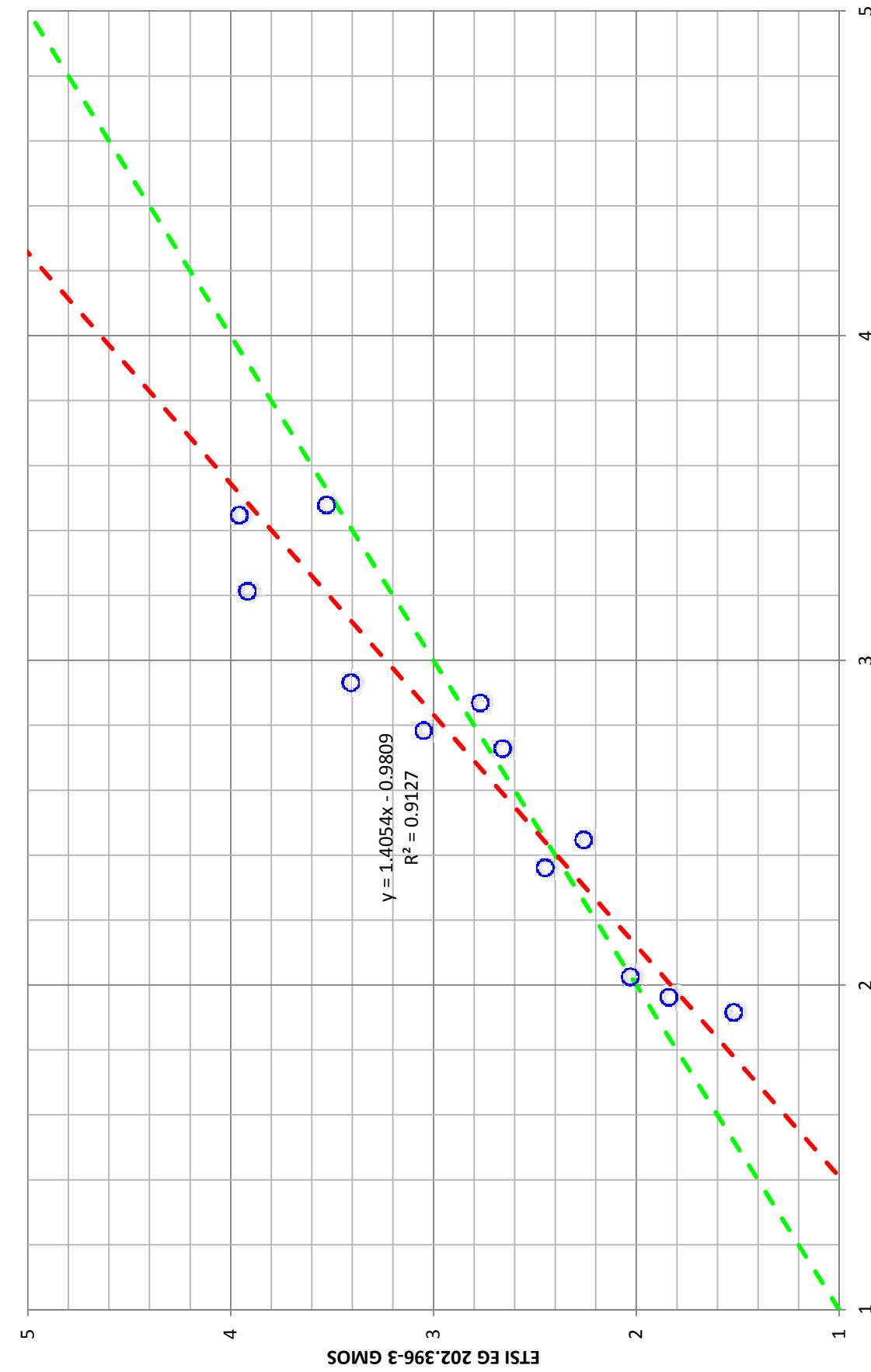
10/24/2011

Tdoc S4-AHQ034

P.835 BAK

13

Qualcomm P.835 MOS x ETSI EG 202.396-3 G-MOS (reference set only)



10/24/2011

Tdoc S4-AHQ034
P.835 MOS

14

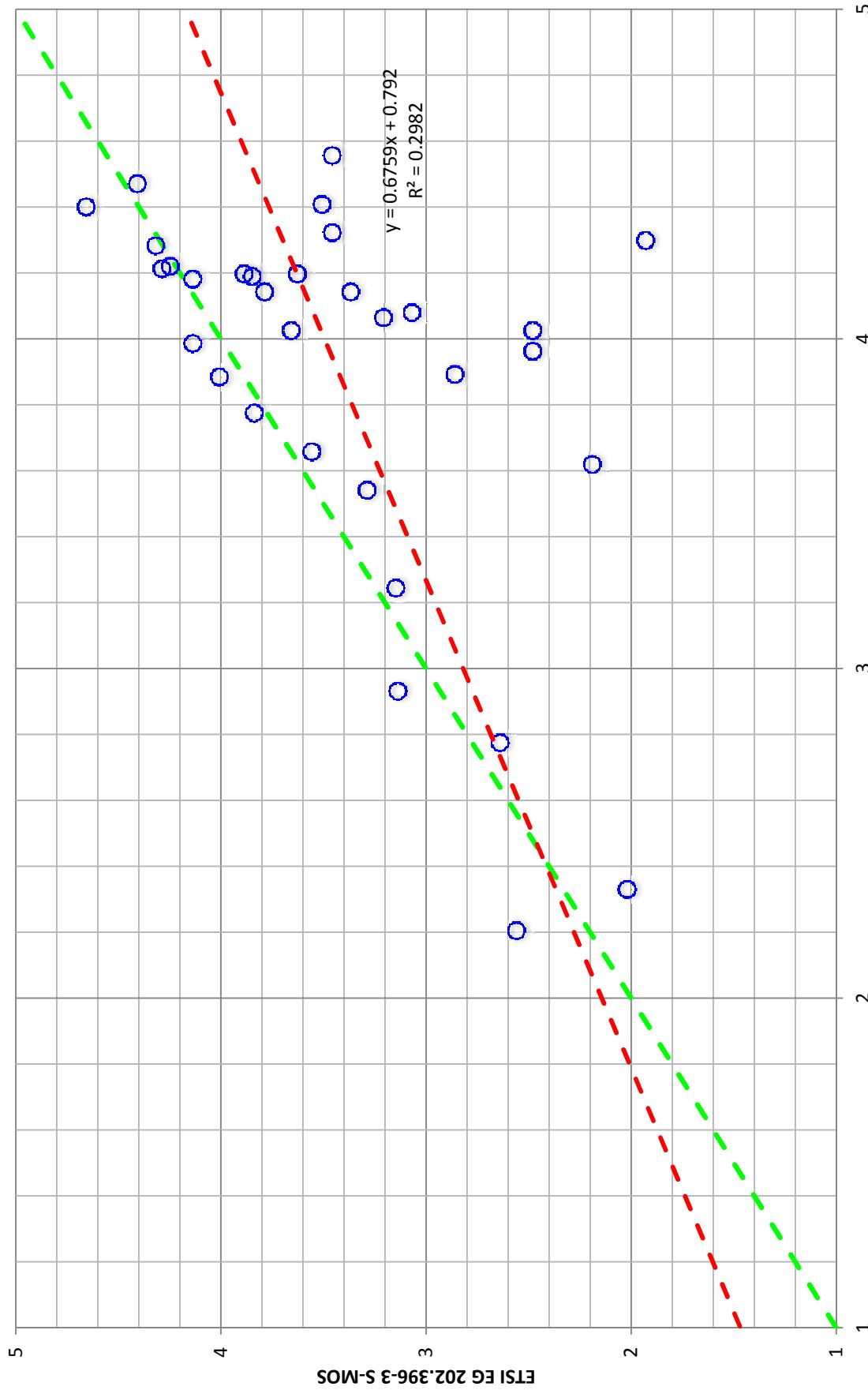
Preliminary observations

- The same trends as observed in the scatter plots between the two subjective tests were observed as expected:
 - Good correlation between subjective and objective scores.
 - BAK scores are lower in Qualcomm LOT
 - Differences in BAK are propagated to overall (MOS) scores.

Correlation of Qualcomm experiment and objective scores (ETSI EG.202.396-3) for all conditions

- The ETSI EG.202.396-3 was applied on the concatenated sentences and the resulting objective score is compared with the subjects mean opinion score.
- The correlation for all conditions is now presented.

Qualcomm LOT P.835 SIG x ETSI EG.202.396-3 S-MOS (all conditions) (R=0.546)

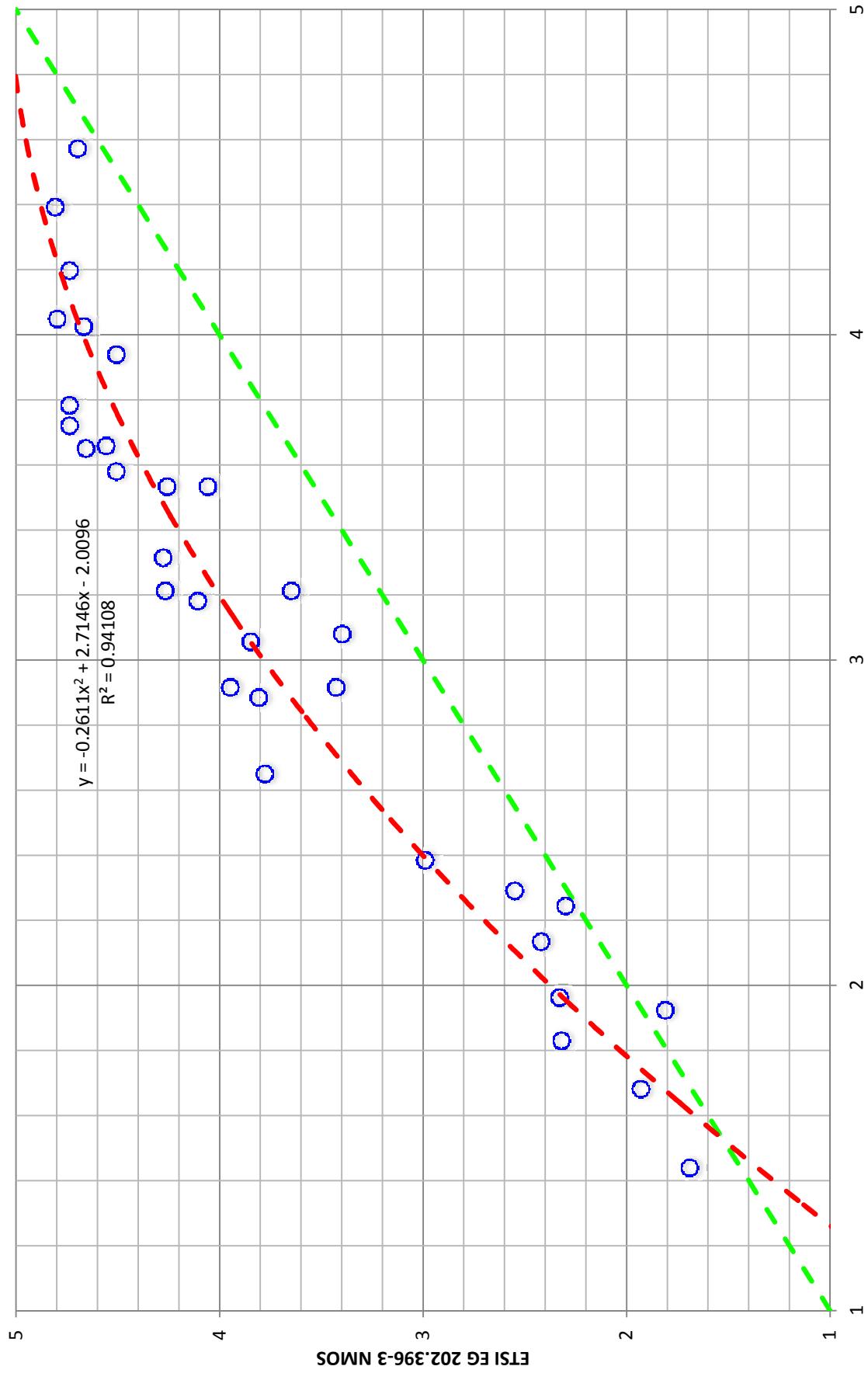


10/24/2011

Qualcomm LOT P.835 SIG
Tdoc S4-AHQ034

17

Qualcomm P.835 BAK x ETSI EG.202.396-3 N-MOS (all conditions) (R=0.970)

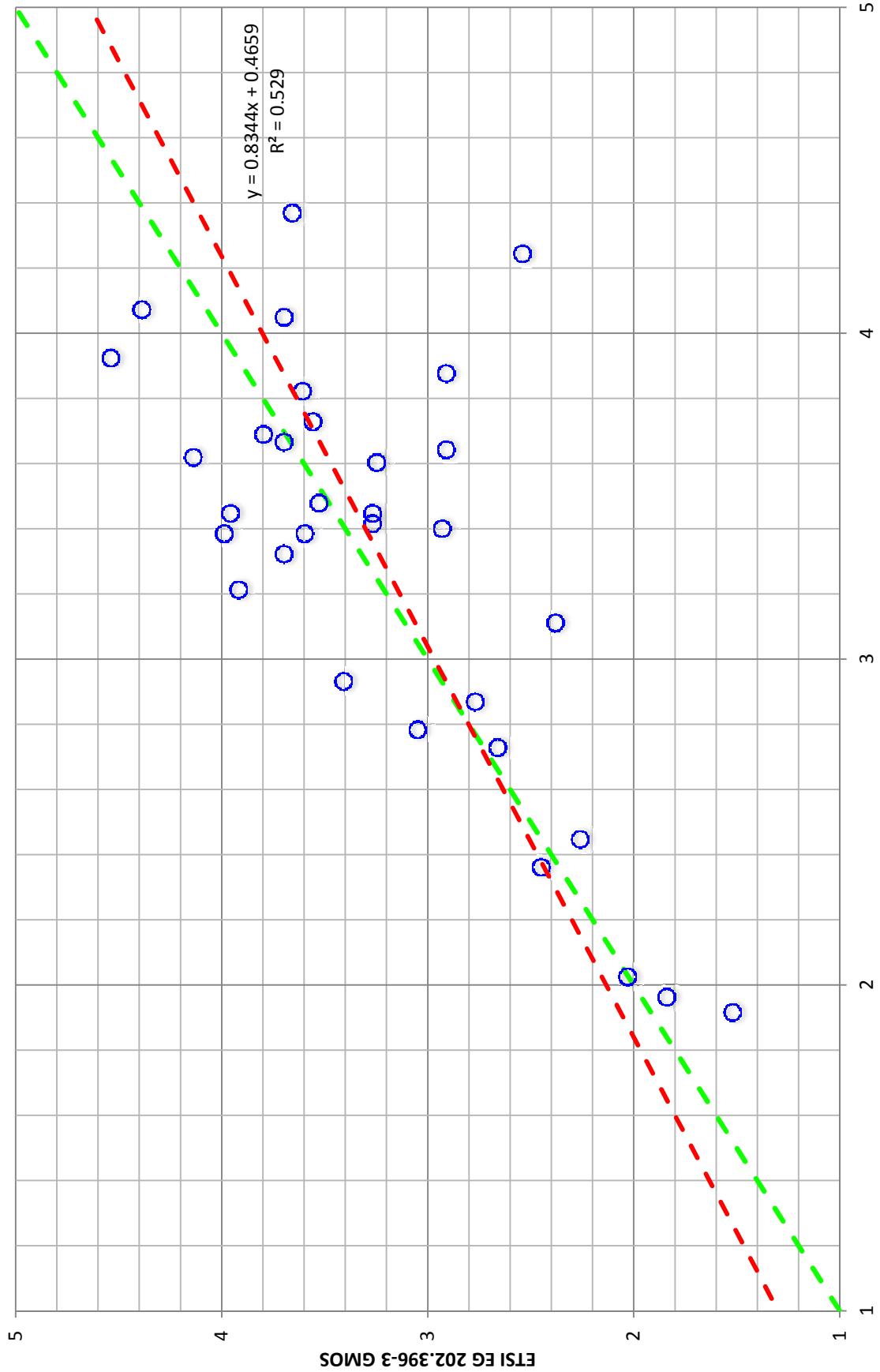


10/24/2011

Qualcomm LOT P.835 BAK
Tdoc S4-AHQ034

18

Qualcomm P.835 MOS x ETSI EG 202.396-3 G-MOS (all conditions) (R=0.727)



10/24/2011

P.835 OVRL (Qualcomm LOT)
Tdoc S4-AHQ034

19

Preliminary observations

- The correlation for BAK and N-MOS is good, although mapping for this experiment would be necessary for absolute score prediction.
- The ETSI EG 202.396-3 method seems to underpredict the results of this test for SIG in the case of more aggressive suppressors.
- The potential underprediction in SIG leads to a low correlation in OVRL (MOS) scores as well.
- Device “E” in particular is substantially underpredicted, causing the correlation to go down.
- Device “E” is the most aggressive suppressor and the highest performer in the naïve participants subjective test.
- Curiously, device “E” has ranked as a poor performer in tests previously conducted with expert listeners. Also, the SIG scores for device “E” appear to be lower than other devices for some of the noise conditions indicating the presence of perceivable signal degradation (although still not statistically significant) for the device.
- It will be interesting to investigate the performance of this same device in tests conducted with a different population (outside US) to investigate on potential cultural and conditioning differences.

Conclusion and proposed next steps (for discussion)

- Preliminary analysis seems to indicate that the ETSI EG.202.396-3 method may underpredict P.835 scores in the case of very high noise suppression devices.
- It is however unclear how these devices would perform with a different population group.
- It is suggested to repeat the exact same test in a different country to establish repeatability of P.835 results among different population groups (to be presented at SA4#66(?)).
- Discuss an improved and commonly agreed reference system for subjective tests that includes a clean condition to anchor the extreme of the scale (wait for LS reply from Q7/12).
- If further training tests for the model are felt needed, reduce the confidence intervals of the training database (current ETSI database has only 24 votes per condition, suggest > 128) and include dual microphone noise cancellation terminals in the mix of training conditions.
- Investigate any dependencies of the ETSI model on speech material (ETSI).