

Authors' Reply²

The authors would like to respond to the comments of the reader. First some remarks on the two assumptions, both of which are indeed used in the statistical analysis.

Assumption 1: "There is a one-dimensional latent (non-measured) variable that denotes the perceived quality of the audio systems."

Assumption 1 means that the perceived difference in sound quality of the two systems (namely, systems that differ in sampling rate) can be statistically modeled by a one-dimensional latent variable. Indeed, this assumption should preferably be supported with experimental evidence, but such evidence does not exist. On the other hand, the assumed model was tested for goodness-of-fit (this is not included in the paper) and it was not rejected. If the latent variable was in fact multidimensional, this may have led to a rejection of the model. In case of a multidimensional latent variable, the assumed one-dimensionality is of course at best an approximation to reality.

Assumption 2: "Both digital systems X and Y have a performance that is not better than the performance of the analog system A, if we measure this performance on the one-dimensional latent variable."

Assumption 2 has reasonable logic from audio knowledge and experience if we assume that in theory the analog version has an infinite number of data points (that is, infinite resolution), and DXD (8 Fs) has used eight times more data points for conversion than the 1-Fs system.

The author of the comment is correct in stating that there is an indication that in our listening experiment "the cues involved are of a different nature for different people." In assessing very small sonic differences, such as in this experiment, this indeed must be the case. However, he is not correct in suggesting that this absolutely rejects the existence of a "one-dimensional latent (nonmeasured) variable that denotes the perceived quality of the audio system" required by the statistical model.

Each subject was asked in the questionnaire to choose which of the two digital paths sounds closer to the analog reference *in every aspect of sound quality they can find*. We asked for one cumulative assessment that integrates all aspects of sound quality into one judgment. Any individual differences in focus, point of attention, ability to discriminate, relative balance in weighting different aspects of quality, and so on, should be transparent in the final vote required for "which of the two digital paths sounds the closest to the analog reference in every detail of sound quality." The fact that "these characteristics are not known, neither is their number" does not reject the existence of a one-dimensional latent (nonmeasured) variable.

Further, the phrase "In this test, we have made a general assumption that the high-sampling version was closer in quality to the reference system than the low-sampling version, simply on the basis of it having 8 times greater number of data points in the conversion from analog, and without regard for the perceived difference between these

digital formats" seems to have led to some misunderstanding. It is important to realize that this assumption has not been used as prior knowledge in testing and in the statistical analysis of the data, so conclusions have not been biased by this hypothesis. The phrase should be considered as a hypothesis made in advance of the study that could be verified by the experiment. The subjects were not informed of this assumption and it served only as our own reference for giving a 1 or a 0 to each response in a consistent manner for subsequent processing. The subjects were given no hints in the questionnaire as to what would be the difference between the two digital systems. This assumption was therefore not used in any way to determine or affect the outcome of the test.

In bandwidth condition C2, subject 6 indeed has a deviating behavior as compared to the other subjects. When testing the hypothesis $p = 0.5$ per subject (p as defined in the paper) this is the only subject supplying a result that significantly deviates from the hypothesis $p = 0.5$. Note, however, that testing the hypothesis $p = 0.5$ per subject seems not very effective. With six trials only the extreme outcomes of 0 and 6 will be statistically significant, and the testing power is not high. Note that all subjects could have a score of 5 and there would be no statistical significance. It seems better to test the hypothesis $p = 0.5$ by combining all subject results. Furthermore it looks a little rushed to reject Assumption 1 just based on the outcome of the single subject 6. As stated before, the model goodness-of-fit test was not significant. Also, this subject may be one who did not understand the task and may therefore even have to be considered as exceptional.

We performed a statistical test without making Assumptions 1 and 2, for each of the two conditions separately. We tested the null hypothesis that subjects have no preference for system X or Y in matching the quality of the analog system versus the alternative hypothesis that there is a preference for system X or Y (two-sided test). With the exact binomial test and when testing at the 5% level of significance we found an almost significant result, with significance probability 5% for condition C1, and no significant result for condition C2.

One needs to be clear about the limitations of this experiment. Two separate tests were conducted, each providing a different reproduction bandwidth condition, C1 (100 kHz) and C2 (20 kHz) for a low-Fs and a high-Fs audio signal, respectively. Listeners could not directly compare bandwidth conditions C1 and C2 for the high-Fs signal, the low-Fs signal, or the analog reference. A purpose-built mechanical sound source and anechoic acoustic environment were used in the experiment to make the live comparison to the analog reference possible. Deriving conclusions that are universal and depart substantially from these conditions is a matter of interpretation. If the results seem counterintuitive, additional experiments should be conducted.

Our conference publication is an engineering report describing an experiment employing a live sound source as an analog reference. We presented our conclusions, but we also provided the data for anyone else to process, should they wish to do so, to draw their own conclusions.

²Manuscript received 2009 March 24.

We welcome anyone to repeat the experiment based on whatever changes are deemed appropriate.

Considering that the experience of professional recording and mastering engineers working with high-resolution audio has not been confirmed and quantified in laboratory tests, we do not have a good validation of the testing methods used in our industry. The current subjective test-

ing methodology does not seem to sufficiently reveal or amplify the features characterizing individual listening experiences. Perhaps this methodology, which is derived from food and fragrance testing, is not as readily effective for investigating subtle auditory sensations and the experience of music? We too would like to encourage more work in this area.

WIESLAW WOSZCZYK, *AES Fellow*
Schulich School of Music of McGill University
Montreal, QC, Canada H3A 1E3

JAN ENGEL
Centre for Quantitative Methods COM BV
5611 BK Eindhoven, The Netherlands

JOHN USHER³
Schulich School of Music of McGill University
Montreal, QC, Canada H3A 1E3

RONALD AARTS, *AES Fellow*
Philips Research
5656AA Eindhoven, The Netherlands

DERK REEFMAN, *AES Associate Member*
Philips Research
5656 AA Eindhoven, The Netherlands

³Now with Barcelona Media, 08018 Barcelona, Spain.