Audio Engineering Society

Convention Paper

Presented at the 118th Convention
2005 May 28–31 Barcelona, Spain

Selecting participants for listening tests of multichannel reproduced sound

Florian Wickelmaier[1], and Sylvain Choisel[1,2]

[1] *Sound Quality Research Unit, Department of Acoustics, Aalborg University, 9220 Aalborg, Denmark*

[2] *Bang & Olufsen A/S, 7600 Struer, Denmark*

Correspondence should be addressed to Florian Wickelmaier (fw@acoustics.dk)

ABSTRACT

A selection procedure was devised in order to select listeners for experiments in which their main task will be to judge multichannel reproduced sound. 91 participants filled in a web-based questionnaire. 78 of them took part in an assessment of their hearing thresholds, their spatial hearing, and their verbal production abilities. The listeners displayed large individual differences in their performance. 40 subjects were selected based on the test results. Self-assessed listening habits and experience as obtained from the web questionnaire did not predict the results of the selection procedure. Further, absolute hearing thresholds did not correlate with the spatial-hearing test. This leads to the conclusion that task-specific performance tests might be the preferable means of selecting a listening panel.

1. INTRODUCTION

In many experiments involving human listeners the experience or expertise of the participants is of crucial importance. The experimenter has to make a decision as to whether the subjects should be naïve (unexperienced with respect to the task) or experts. Clearly, if generalizability of the experimental results was the only concern, one would randomly sample the level of experience rather than restrict the sample to only that, potentially small, part of the population which can be regarded as "expert listeners" by any given criterion. On the other hand, human behavior is always characterized by an intrinsic random component, which often makes it a difficult task to extract the systematic effects, unless certain sources of variation have been eliminated a priori. It is often assumed, and there is empirical evidence [1], that expert listeners display less variation in their judgments and are therefore more reliable.

One way of dealing with this dilemma between generalizability and reliability is to select the participants randomly and subsequently train them to become experts [2]. While this strategy might be applicable to many experiments, great care has to be taken in order not to bias subjects' judgments by the training procedure. The risk of biasing listeners, however, is particularly high in studies having an exploratory character because subjects might be asked to make judgments about a variety of auditory sensations, possibly even new and as yet unlabeled ones. Since the present selection was made for such a study, training subjects was disregarded. Rather the strategy in the present study was to start from a random sample of participants and select the best ones according to specified criteria. In contrast to other procedures (e. g. the Generalized Listener Selection (GLS) procedure [3]), which are similar in spirit, but base the selection on general listening abilities, the current procedure presents the participants with specific tests which are related to the abilities required in later tasks.

The participants were selected for a series of experiments which aim at uncovering auditory attributes of multichannel reproduced sound. It will be important that the panelists can appreciate the differences between different reproduction modes (like mono, stereo, 4- and 5-channel surround, etc.). It will also be important that they possess good verbal abilities, especially when it comes to promptly producing a description of their sensations. The challenge for the selection procedure is to assess these abilities without telling the subjects what to listen for or what to describe, and thereby irrevocably biasing their judgments. For the two desired abilities this problem was addressed in the following ways: A discrimination test of sounds varying in stereo width was conducted employing a three-interval forced-choice procedure, which circumvents naming the involved attribute both by the experimenter and the subject. It was assumed that listeners with better discrimination could also differentiate the reproduction modes more easily in later experiments. The verbal production abilities were assessed via a standard verbal fluency test [4], assuming that participants with a high fluency score can describe their sensations more readily.

It is generally recommended (e. g. by the International Electrotechnical Commission (IEC) [5]) that an audiometric test should not be the (only) means for selecting a listening panel. In this study, audiometry was used to assess normal hearing of the participants and to supplement the specific ability tests. As a further supplement, data about the listening habits and prior experience of the participants were collected by means of a questionnaire made available via the Internet.

2. METHOD

The selection was conducted in four steps. First, the candidates signed up for the tests by filling in a web questionnaire. Invitations to the experiments had been placed on the Internet, as well as in public places, such as libraries, cafeterias, music shops, pubs, and shopping centers. The requirements for participation in the study were (a) to be a native Danish speaker and (b) to be available for the duration of the project (about ten months). After signing up, the participants were invited to the tests proper, which included audiometry, a spatial hearing test, and a verbal fluency test. These tests were conducted in a double-walled sound-insulated chamber.

2.1. Web questionnaire

A web-based questionnaire, inspired by the one used by Mattila & Zacharov [3], was used for registering the subjects' demographic variables and listening experience into a database. The entire questionnaire can be seen in the Appendix. Two questions were used to screen the participants for clinically relevant hearing problems or hearing damage. 91 persons filled in the questionnaire, of which four did not fulfill the language requirement, two were participants in parallel (and possibly biasing) experiments, and seven dropped out. The remaining 78 participated in the selection tests, none of them reported any known hearing problems or damage.

2.2. Audiometry

The next requirement for the 78 listeners was a maximum hearing threshold of 20 dB within 250 Hz to 8 kHz. The audiometric test [6] was performed using a Madsen (model OB 40) audiometer. Twenty of the subjects had already participated in earlier experiments, and recent audiometric data were available.

2.3. Stereo-width discrimination

The second test concerned the subjects' ability to discriminate between sounds which varied in stereo width. Stereo width was manipulated by decomposing the signal into a weighted sum of the sum $(L+R)$ and the difference $(L-R)$ of the left and right channels (Equation 1). This weighted sum is sometimes called the mid/side (MS) ratio, especially when the sound has been recorded with both an omnidirectional/cardioid and a bidirectional microphone. From an original stereo recording with left and right channel L and R, a new sound (L', R') varying in stereo width can be derived by

$$\begin{aligned} L' &= (1-\tfrac{\beta}{2})(L+R) + \tfrac{\beta}{2}(L-R), \\ R' &= (1-\tfrac{\beta}{2})(L+R) - \tfrac{\beta}{2}(L-R), \end{aligned} \qquad (1)$$

where the parameter β determines the stereo width: When β equals one, the left and right channel of the derived sound are identical to the original stereo channels; when β equals zero, L' and R' both amount to the sum of the stereo channels, i. e. mono. By varying β between zero and one, it is possible to create sounds having a different degree of stereo width, from mono to stereo.

2.3.1. Apparatus

A personal computer equipped with a sound card (RME Hammerfall HDSP) connected to an external D/A converter (RME ADI-8 DS) was used to play back the sounds in the MS-ratio test. The stimuli were of approximately 1.5 s duration and were presented over headphones (Beyerdynamic DT990), delivered by a headphone amplifier (Behringer Powerplay 4400). The stereo recording was presented at an A-weighted equivalent level of 76.4 dB SPL to the left and 78.8 dB SPL to the right ear as measured with an artificial ear (Brüel & Kjær 4153). The participants entered their responses by clicking one of three buttons presented on a computer screen.

2.3.2. Procedure

An adaptive procedure (3AFC, 2-up/1-down) [7, 8] was employed in order to assess the reduction in stereo width that was detected 71% of the time. A stereo recording of a piano chord ([9], track 39, at 1'53) served as a standard, and a comparison was derived from it by changing the MS ratio according to Equation 1. The participants performed a forced-choice oddity task. On each trial, they had to identify which of the three sounds was *different* from the other two. According to the subject's response, the comparison varied adaptively from mono ($\beta = 0$) towards stereo ($\beta = 1$), converging on the 71%-discrimination threshold. After responding correctly in two successive trials, β was increased; it was decreased after every wrong answer. The step size decreased with increasing β-value by $0.3(1-\beta)$. Thus, as the task became harder, the step size became smaller, and the upper bound of $\beta = 1$ could never be exceeded. The procedure stopped after eight reversals, and the discrimination threshold was estimated from the average β-value at the last four reversals. In order for the procedure to be less transparent to the listeners, each trial contained either two standards and one comparison, or one standard and two (identical) comparisons, in random order (i. e. the odd sound was either standard or comparison). The subjects were not told what kind of differences to listen for, and therefore were free to use any criterion. In order to remove loudness cues as much as possible, the mono stimulus was attenuated by 3 dB. This attenuation was decreased gradually to zero, as β approached a value of one. Figure 1 displays the stimulus generation schematically.

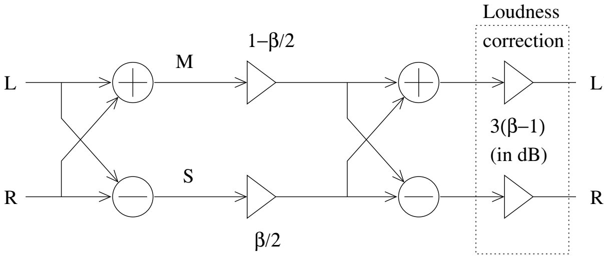


Figure 1: Stimulus generation in the MS-ratio test: The output is a mono signal when $\beta = 0$, and the original stereo when $\beta = 1$. The final stage applies a gain varying between -3 dB (mono) and 0 dB (stereo).

In a pilot experiment including four consecutive measurements of seven listeners, it was observed that the discrimination thresholds improved after the first measurement, and remained constant afterwards. Therefore, before the actual test, the subjects underwent a short familiarization, in which the procedure stopped after two reversals. The results from the familiarization were not incorporated into the threshold estimate. For none of the participants did the measurement last longer than 15 minutes, including familiarization.

2.4. Verbal fluency

In the last test, the subjects' verbal production abilities were assessed in an alternating verbal fluency test [4]. At the beginning of the test the following written instruction (in Danish) was handed out:

> In this task you should within one minute name as many different Danish words as possible which belong alternately to the categories "animals" and "fruits". First name an animal then a fruit, then again an animal, etc. Please do not repeat a word you have already said before.
>
> Please try to say as fast as possible as many different words as you can. Start with an animal.

The participants were seated in the listening booth in which a microphone, connected to the computer in the control room, was used to record the word list. The sound files were saved, and were analyzed at a later point in time. A fluency score was assigned to each word list by counting the correct responses. Incorrect were words not belonging to the categories "animals" or "fruits", newly created words, proper nouns or given names, word repetitions, and category perseverations (naming, e. g., two animals in a row, as in "mouse – goat – nut"). A familiarization session preceded the test, which was identical, but had the two different semantic categories "professions" and "items which can be found in a supermarket". After having checked that the instructions were understood, the familiarization ended. Results from the familiarization were not included in the fluency score.

3. RESULTS

Figure 2 displays the data obtained in the selection procedure. On the abscissa the stereo-width discrimination thresholds are displayed, on the ordinate the fluency test scores. The discrimination thresholds ranged from 0.15 to 0.83 (corresponding to MS ratios between 93:7 and 59:41) with a mean of 0.50 and a standard deviation of 0.18. The fluency scores ranged from 11 to 29 with a mean of 16.6 and a standard deviation of 3.4. Eight out of the 78 participants (marked by open circles) had a mild hearing loss of between 25 and 40 dB in either ear at at least one of the audiometric frequencies between 250 and 8000 Hz, and were therefore rejected.

A further criterion for a-priori rejection was a stereo-width discrimination threshold below chance level. The chance level was determined by means of a Monte-Carlo simulation, in which the outcome of the adaptive procedure was recorded when a virtual subject responded randomly. On each simulation run, 1000 simulated thresholds were generated. Figure 3 shows a typical example of the distribution of the resulting simulated thresholds. The median of this distribution is close to zero (0.08). The chance level was adopted as the 95% percentile of the distribution, which lies at 0.4. In order to take variations due to sampling into account, 100 simulation runs were performed and each time the 95% percentile was estimated. The 95% percentile of the 100 estimates again was found to be 0.4 and consequently set to the criterion of chance performance. The 24 participants having a lower sensitivity than 0.4 were excluded, because it cannot be ruled out that they were guessing while performing the discrimination

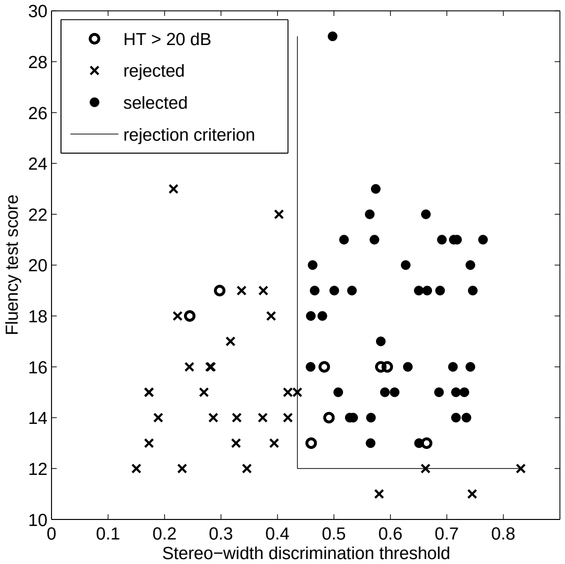


Figure 2: Results of the selection procedure. Listeners above and to the right of the rejection criterion were selected on the basis of the results in the discrimination and fluency test (solid circles). Participants with a hearing threshold (HT) of more than 20 dB were rejected a priori (open circles).

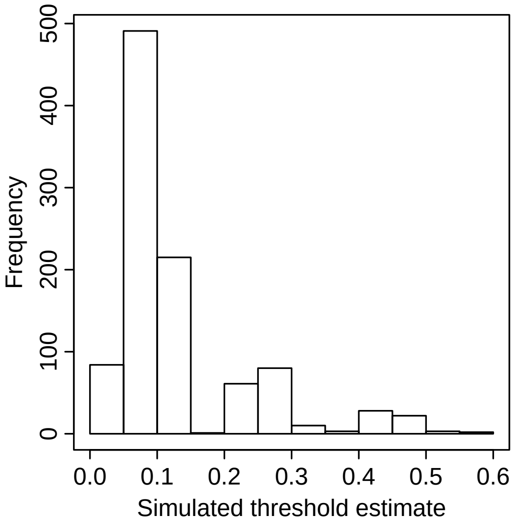


Figure 3: Typical result of the Monte-Carlo simulation of discrimination thresholds, when the responses were given at random. Displayed are the frequencies of 1000 simulated thresholds.

task.

In order to select the final listening panel, the following decision rule was applied to the remaining subjects: A subject was removed from the list of candidates, if he or she performed worst at either the stereo-width discrimination or the verbal fluency task. In doing so, both tasks were weighted equally. This elimination process stopped when 40 subjects were left. The selected listeners lie in the upper right quadrant of Figure 2 (marked by solid circles). They are separated from the rejected subjects (crosses) by a horizontal and a vertical line. These lines correspond to rejecting the worst cases both with respect to verbal production abilities and sensitivity to changes in stereo width. Another eight listeners were excluded on the basis of this criterion over and above those previously excluded due to hearing loss, or below-chance performance in the spatial hearing test. The remaining 40 subjects were selected for participating in later listening experiments.

3.1. Stereo-width discrimination and demographic variables

The influence of the demographic variables acquired via the web questionnaire (see Appendix) on the sensitivity to stereo width was investigated. In particular, sex, occupational background of the participants, habits concerning listening to music, attending concerts, or going to the cinema, playing an instrument, owning a hi-fi or a surround sound system, considering oneself as a critical listener, and being professionally involved in music or audio were included in the analysis. Frequently, such variables are considered potential predictors of listening abilities. Table 1 shows the average discrimination thresholds stratified by those variables together with the sample size and standard deviation.

Contrary to the expectations, however, only the variable sex turned out to have a significant influence on the discrimination threshold, as confirmed by a two-sample t-test [$t(66) = 3.66$; $p < .001$]. On average the male subjects were by about 0.9 of a standard deviation more sensitive than the females. In order to investigate interactions with the occupational background of the participants, they were assigned according to their profession the five categories engineering, languages and communication, music and music therapy, social sciences, and others. The majority of the participants were students of the respective fields. A two-factor analysis of variance revealed no significant interaction between background and discrimination threshold [$F(4, 60) = 1.71$; $p = .159$], nor a significant main effect of the background [$F(4, 60) = 2.31$; $p = .068$], but a highly significant gender effect [$F(1, 60) = 12.94$; $p < .001$]. A similar result (main effect of sex only) was obtained when analyzing an interaction with the self-assessed professional experience. Since there was no reason to a priori expect better performance of the male participants in the stereo-width discrimination tests, no gender correction was applied to the tests results. As a consequence, proportionally more females than males were rejected based on this criterion.

3.2. Semantic fluency and demographic variables

No significant differences between groups based on sex, education and age were found in semantic fluency. Table 2 shows the mean fluency scores, standard deviations and sample sizes stratified by sex, years of education and age. A three-factor analysis of variance revealed no significant influence of the interaction of sex, education and age [$F(1, 65) = 0.72$; $p = .398$], nor significant two-way interactions or main effects. The variation of the fluency scores can

Table 1: Estimated stereo-width discrimination thresholds stratified by demographic variables, self-assessed experience, and listening habits. Only the difference between males and females is statistically significant. Note—n sample size, M mean, SD standard deviation.

		Est. threshold	
Category	n	M	(SD)
Sex			
male	43	0.56	(0.18)
female	27	0.41	(0.16)
Background			
music	13	0.59	(0.15)
engineering	27	0.51	(0.18)
languages	9	0.51	(0.22)
social science	15	0.42	(0.17)
others	6	0.45	(0.20)
Professional experience			
yes	20	0.56	(0.17)
no	50	0.48	(0.19)
Listening to music			
daily	59	0.50	(0.19)
weekly	11	0.50	(0.18)
Attending concerts			
weekly/monthly	32	0.54	(0.18)
rarely or not	38	0.47	(0.18)
Playing instrument			
daily	25	0.55	(0.20)
rarely or not	45	0.48	(0.17)
Critical listener			
yes	63	0.50	(0.18)
no	7	0.49	(0.22)
Going to cinema			
monthly	43	0.49	(0.20)
less than monthly	27	0.53	(0.16)
Own hi-fi system			
yes	54	0.52	(0.18)
no	16	0.45	(0.19)
Own surround system			
yes	14	0.44	(0.18)
no	56	0.52	(0.18)
Participated in tests			
yes	25	0.48	(0.17)
no	45	0.52	(0.19)

therefore be attributed to the individual differences in the sample.

Sample percentiles of the 78 native Danish speakers are displayed in Table 3. Participants having a score of less than 13 were rejected according to the rejection criterion (cf. Figure 2). This corresponds to excluding the subjects in the lower 10% of the distribution.

Table 2: Semantic fluency scores (animals–fruits) stratified by sex, years of education and age. Note—n sample size, M mean, SD standard deviation.

		Fluency score	
Category	n	M	(SD)
Sex			
female	28	17.1	(3.1)
male	50	16.3	(3.6)
Education, years			
less than 13	5	15.8	(3.6)
13–16	36	17.1	(3.8)
more than 16	37	16.2	(3.0)
Age, years			
20–24	41	16.8	(3.8)
25–29	29	16.5	(3.1)
30–44	8	15.6	(2.2)

Table 3: Sample percentiles of the semantic fluency test.

Percentile	10	25	50	75	90
Fluency score	13	14	16	19	21

4. DISCUSSION

The participants displayed considerable variation in the results of the specific ability tests. Therefore, these two tests are especially suited for selection purposes. The web-based questionnaire, however, failed to explain the differences in sensitivity to changes in stereo width entirely. Especially, the questions related to (self-assessed) prior experience or to listening habits of the subjects provide no good means for predicting the results. From these findings it might be concluded that such investigations into the *attitudes* of potential panelists should have little priority for their selection, whereas the main focus should be put on their *behavior* observed in specific tests.

The fluency scores, both in terms of mean value and standard deviation, are comparable with results from studies of native English speakers [10, 11, 12,

Table 4: Results of alternating word fluency tests among native English speakers. The bottom line shows the present sample (native Danish speakers). Note—n sample size, M mean, SD standard deviation, age and education in years.

		Fluency score		Age		Education	
Category	n	M	(SD)	M	(SD)	M	(SD)
Boy's names–fruits	20	17.1	(4.0)	63.5	(10.1)	9.9	(3.5)
M-words–vegetables	60	14.5	(3.0)	29.1	(6.4)	–	–
Fruits–furniture	11	16.0	(3.3)	68.1	–	14.6	–
L-words–R-words	9	9.7	(2.5)	54.8	(8.2)	11.7	(2.8)
Colors–occupations	45	14.0	(3.8)	63.1	(10.6)	13.6	(3.1)
Animals–states	45	17.5	(5.4)	63.1	(10.6)	13.6	(3.1)
C-words–P-words	45	10.2	(4.6)	63.1	(10.6)	13.6	(3.1)
Animals–fruits	78	16.6	(3.4)	25.8	(5.0)	–	–

13, 14]. The comparative results can be seen in Table 4. Note, however, that neither the age structure nor the semantic (or lexical) categories exactly match those of the present study.

Finally, as expected, it was found that the discrimination thresholds in the spatial hearing test cannot be predicted by the absolute thresholds measured in the audiometry. The correlation between the maximum hearing threshold per subject obtained at any of the frequencies at either ear with the spatial discrimination threshold was not significant [$r = .19$; $p = .115$]. Therefore, when the panelists' task is to judge supra-threshold stimuli, the selection should not be determined by audiometry but rather by performance at specific tests which are related to the later requirements in the panel.

Concluding remarks

The major assumption underlying the selection procedure is that the selected listeners would outperform the non-selected ones in later experiments, be it by their superior ability to discriminate the sounds, by their better verbalization skills, or generally by an increased reliability of their judgments. In the present study, however, no attempt was made to validate this assumption against an external criterion. This is left to other researchers who might find the procedure interesting and helpful for their own studies of multichannel reproduced sound. The advantages of the two tests chosen are, that they allow for an efficient assessment of listeners, give rise to sufficient variance between them, and are easily analyzed and reported in quantitative indices.

5. ACKNOWLEDGMENTS

This manuscript was completed while the authors were employed at the Sound Quality Research Unit at Aalborg University. This unit receives financial support from Delta Acoustics & Vibration, Brüel & Kjær, and Bang & Olufsen, as well as from the Danish National Agency for Industry and Trade (EFS) and the Danish Technical Research Council (STVF). The authors would like to thank Wolfgang Ellermeier for his helpful comments on an earlier draft of the manuscript.

6. REFERENCES

[1] S. Bech, “Selection and training of subjects for listening tests on sound-reproducing equipment,” *Journal of the Audio Engineering Society*, 40:590–610, 1992.

[2] S. Bech, “Training of subjects for auditory experiments,” *Acta acustica*, 1:89–99, 1993.

[3] V. V. Mattila and N. Zacharov, “GLS – A generalised listener selection procedure,” *Journal of the Audio Engineering Society (Abstracts)*, 49:546, 2001.

[4] O. Spreen and E. Strauss. *A compendium of neuropsychological tests.* Oxford University Press, New York, 1998.

[5] CEI IEC technical report 60268-13. Sound system equipment – Part 13: Listening tests on loudspeakers. International Electrotechnical Commission, 1998.

[6] ISO 389-1985 (E). Standard reference zero for the calibration of pure tone air conduction audiometers. ISO, Geneva, 1985.

[7] H. Levitt, "Transformed up-down methods in psychoacoustics," *Journal of the Acoustical Society of America*, 49:467–477, 1971.

[8] W. Jesteadt, "An adaptive procedure for subjective judgments," *Perception & Psychophysics*, 28:85–88, 1980.

[9] EBU document Tech. 3253. Sound quality assessment material. Recordings for subjective tests. Users' handbook for the EBU-SQAM compact disc. European Broadcasting Union, 1988.

[10] R. Zec, E. Landreth, J. Belman, S. Fritz, A. Hasara, W. Fraiser, S. Wainman, M. McCool, E. Grames, C. O'Connell, R. Harris, R. Robbs, R. Elble, and B. Manyam, "A Comparison of Phonemic, Semantic, and Alternating Word Fluency in Parkinson's Disease," *Archives of Clinical Neuropsychology*, 14:255–264, 1999.

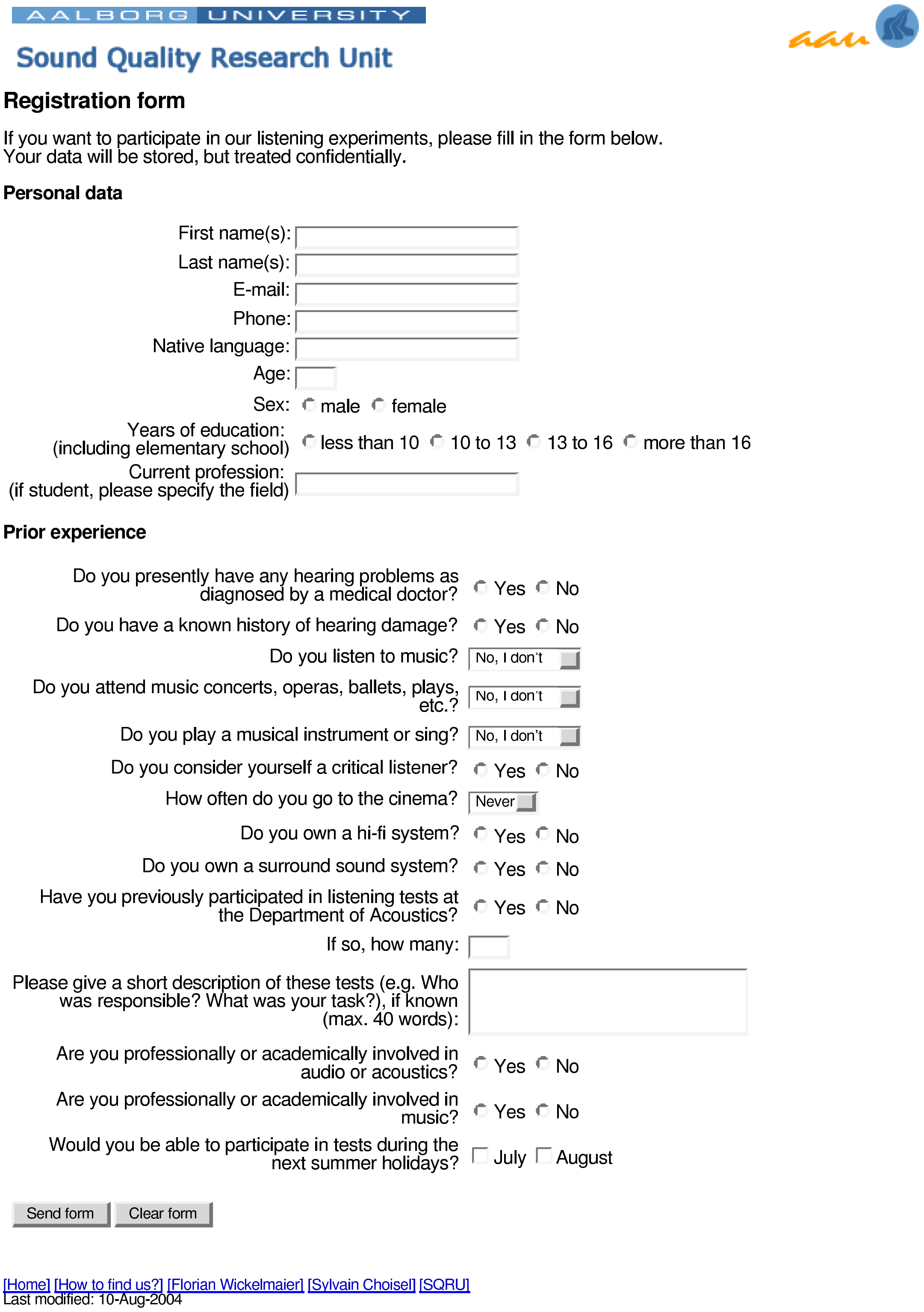
[11] L. H. Phillips, R. Bull, E. Adams, and L. Fraser, "Positive Mood and Executive Function Evidence from Stroop and Fluency Tasks," *Emotion*, 2:12–22, 2002.

[12] J. V. Baldo, A. P. Shimamura, D. C. Delis, J. Kramer, and E. Kaplan, "Verbal and design fluency in patients with frontal lobe lesions," *Journal of the International Neuropsychological Society*, 7:586–596, 2001.

[13] K. Dujardin, L. Defebvre, P. Krystkowiak, S. Blond, and A. Destée, "Influence of chronic bilateral stimulation of the subthalamic nucleus on cognitive function in Parkinson's disease," *Journal of Neurology*, 248:603–611, 2001.

[14] C. A. Bouquet, V. Bonnaud, and R. Gil, "Investigation of Supervisory Attentional System Functions in Patients With Parkinson's Disease Using the Hayling Task," *Journal of Clinical and Experimental Neuropsychology*, 25:751–760, 2003.

APPENDIX: WEB-BASED QUESTIONNAIRE



Sound Quality Research Unit

Registration form

If you want to participate in our listening experiments, please fill in the form below.
Your data will be stored, but treated confidentially.

Personal data

First name(s):

Last name(s):

E-mail:

Phone:

Native language:

Age:

Sex: ○ male ○ female

Years of education: (including elementary school) ○ less than 10 ○ 10 to 13 ○ 13 to 16 ○ more than 16

Current profession: (if student, please specify the field)

Prior experience

Do you presently have any hearing problems as diagnosed by a medical doctor? ○ Yes ○ No

Do you have a known history of hearing damage? ○ Yes ○ No

Do you listen to music? No, I don't

Do you attend music concerts, operas, ballets, plays, etc.? No, I don't

Do you play a musical instrument or sing? No, I don't

Do you consider yourself a critical listener? ○ Yes ○ No

How often do you go to the cinema? Never

Do you own a hi-fi system? ○ Yes ○ No

Do you own a surround sound system? ○ Yes ○ No

Have you previously participated in listening tests at the Department of Acoustics? ○ Yes ○ No

If so, how many:

Please give a short description of these tests (e.g. Who was responsible? What was your task?), if known (max. 40 words):

Are you professionally or academically involved in audio or acoustics? ○ Yes ○ No

Are you professionally or academically involved in music? ○ Yes ○ No

Would you be able to participate in tests during the next summer holidays? □ July □ August

[Home] [How to find us?] [Florian Wickelmaier] [Sylvain Choisel] [SQRU]
Last modified: 10-Aug-2004