Jan BERG
School of Music, Luleå University of Technology, Sweden
The Interactive Institute, Piteå, Sweden

Evaluation of perceived spatial quality of 5-channel microphone techniques by using selected spatial attributes[1]

«Subjektive» Qualitätsbewertung von 5-Kanal Mikrofontechniken mit Hilfe ausgewählter räumlicher Attribute

As the number of techniques and systems for spatial audio rendering increases, evaluation of the perceived spatial quality of reproduced sound is becoming more important. Earlier experiments indicated that the perceived spatial quality consists of a number of perceptual dimensions describable by attributes. These attributes were previously found relevant for describing the spatial quality of stimuli subjected to different modes of reproduction. In this paper, new attributes was elicited and the applicability of these and previously encountered attributes for assessment of spatial quality was tested in the context of new stimuli, recorded by means of 5-channel microphone techniques and reproduced through a 5.0 system. The results showed that the selected spatial attributes enabled a group of listeners to differentiate between the different microphone techniques.

INTRODUCTION

A fundamental question in audio work concerns the audio quality and ways to assess it. The quality of systems for spatial audio becomes more and more interesting as the multichannel techniques for recording, transmission and reproduction of audio develops. Salient features of these techniques are their enhanced ability to enable the listener to perceive the location of sounds and the sense of the acoustical environment in which the sound source is located. This can also be described as the aptitude to detect "the three-dimensional nature of the sound sources and their environment". The performance of a sound system in this respect is denoted as "spatial quality". As it refers to the sensations perceivable by a human listener, spatial quality is a concept in the perceptual domain.

Different processes applied in the audio production chain are likely to affect different properties of the audio signal, including the spatial quality. To be able to evaluate the influence of these processes, methods for detecting and quantifying the audible differences between the processes must be found. One approach is to assess reproduced sounds on a

[1] This paper is submitted by request of the TMT 2002 and contains a condensed version of the paper [18].

holistic basis, i e to evaluate the sound as an entity. As there are other properties of a reproduced sound than the features described by the term spatial quality, there is a risk of confusing spatial and non-spatial properties and also a difficulty in how to weigh these in order to come up with a general assessment of the sound. In an evaluation situation, it is also possible that non-spatial properties have a strong influence on perception, thereby masking spatial features. An obvious example of this is severe harmonic distortion, drawing the listener's attention away from the position of sound sources in a recording. Another approach to evaluation is to dissect the perception of the reproduced sound into the perceivable components or dimensions that constitutes the total perception of the sound, in order to assess these components separately. The knowledge of these components may result in possibilities to manipulate them, or to simply select the components of interest in an analysis.

The author's approach to this is to consider and adapt methods found in psychology for eliciting and structuring information from listeners, describing the perceived features of reproduced sound. Methods possible for this are reviewed by Rumsey [1]. Of particular interest is the Repertory Grid Technique, originally described by Kelly [2] and later refined and applied by authors in different contexts [3, 4, 5]. The method relies on communication of listeners' conceptions in the form of verbal constructs. In this application, the method is used for eliciting the sensations perceived by a listener exposed to reproduced sound. Another example of a technique used for collecting and structuring verbal information, used in food research, is the Quantitative Descriptive Analysis [6]. Development of descriptive language for speech quality in mobile communications has been utilised by Mattila [7], and for spatial sound by Koivuniemi and Zacharov [8]. In recent years, graphical techniques have been suggested and employed by Wenzel [9], Mason et al [10] and Ford et al [11].

In an attempt to find relevant dimensions of spatial quality, an experiment was conducted in 1998 by Berg and Rumsey [12]. The experiment 's approach was to try to elicit information from the participating subjects by playing back a number of reproduced sounds to them, where after they were asked for verbal descriptions of similarities and differences between the sounds. The subjects then graded the different sounds on scales constructed from their own words. This was an example of a technique where the subjects came up with descriptions using their own vocabulary with known meaning to them, instead of being provided with the experimenter's descriptors for the scales. The data was subsequently analysed by methods used in the Repertory Grid Technique, with the intention to find a pattern or a structure not necessarily known to the subjects (or the experimenters) themselves.

The experimental idea was to investigate if a pattern with distinguishable groups of descriptors showed, and if so, it would be regarded as an indicator of the presence of the underlying dimensions searched for. The results from the experiment have been reported in [12,13,14,15], and indicated the existence of a number of dimensions described by attributes generally used by the subjects for describing perceived differences between spatial audio stimuli. In [15] the correlation between different classes of the attributes was reported. Attributes as descriptors for spatial sound features are also employed by Zacharov and Koivuniemi in their work [16].

To, if possible, validate the findings in the analyses of the 1998 experiment, an experiment was designed and completed in 2001 [17]. The experiment comprised a compilation of the previously extracted attributes from which scales were constructed. The scales were provided to a group of subjects that used them for assessing stimuli with differences in the modes of reproduction (mono, phantom mono and 5-channel techniques). The result was that all attributes provided were valid for discriminating between different combinations of the stimuli. In the discussion of the paper reporting on the 2001 experiment, the authors suggested further testing and validation of the method and the attributes by stating: "… the difference between stimuli can be decreased and more precisely controlled. This will make it possible to observe whether the scales depending on certain attributes are still valid under new conditions. These differences could be created in the recording domain, e.g. by means of different microphone techniques, without changing the modes of reproduction."

As a result of the 2001 experiment, a new experiment [18] was designed to find if a new set of stimuli still would give significant results in terms of the attributes' applicability and thereby validate the selected attributes in the context of evaluation of different 5-channel microphone techniques. This experiment seeks to answer basically the same questions as in the 2001 experiment, but now with stimuli recorded with different recording techniques (microphone set-ups) and without differences in modes of reproduction, having potentially smaller and more subtle differences:

- Are these attributes valid for describing the spatial quality of (a subset of) reproduced sounds?
- Are scales defined by words interpreted similarly within a group of subjects?
- If such scales are found to be valid, which attributes are either correlated or non-correlated?

In order to answer these questions, the new experiment started with a pre-elicitation to find new attributes. These were subsequently compared with the attributes previously encountered

in the 2001 experiment and if new attributes were found, they were added to the list of attributes employed in the new experiment. Scales were constructed from the list of attributes and were provided to a partially new group of subjects. The subjects assessed a number of sound stimuli on the provided scales. The hypothesis to be tested in the experiment and its alternative were:

- If the scales are not relevant for describing parts of spatial quality of a subset of reproduced sounds, they will have insufficient common meaning to the subject group, which will not be able to make distinctions between any stimuli at a significant level, i e the data will contain mostly randomly distributed points.
- If, however, the scales are relevant in this respect, the scales will have sufficient common meaning to the group, which will be able to make distinctions between some or all of the stimuli in the experiment at a significant level.

If the alternative hypothesis is true, the interrelations of scales and attributes can be analysed subsequently.

The purpose of the experiment is primarily to investigate if the attributes provided are sufficient for enabling the group of subjects to discriminate between stimuli and to make observations on the attributes' interrelation. The different recording techniques are assumed to create audible differences primarily in the spatial domain, not necessarily encountered in the author's previous experiments. It has to be emphasised that neither an analysis of the properties of the different microphone techniques, nor the physical differences between the stimuli are the primary scope of this paper, although some comments on these will be made.

METHOD

The objective of the experiment was to investigate if a non-naïve group of subjects was able to discriminate in a meaningful fashion between a number of stimuli in the form of recorded sounds on scales defined by certain attributes. The subjects were provided with a list of attributes with associated descriptions. The task was, for every attribute, to listen to a number of different sound stimuli and grade the stimuli on scales defined by the attributes. The list of attributes is a result of analyses of previous experiments, where the applicability of a number of attributes has been tested. In addition to that, before the main experiment reported in this paper commenced, a pre-elicitation experiment comprising a smaller number of subjects was performed. The aim of the pre-elicitation was, for the stimuli selected for the main experiment, to: a) have an indication if the subjects were able to find differences between the

stimuli, and b) elicit attributes describing these differences. The attributes emerging from the pre-elicitation was combined with the previously encountered attributes to form the final list of attributes used in the main experiment. Analyses were made to find if the attributes used enabled the group of subjects to make discriminations between the stimuli and to discover the attributes that were either strongly correlated or independent.

The subjects performed the experiment one at a time in a listening room equipped with loudspeakers and a user interface in the form of a computer screen, a keyboard and a mouse. All communication with the subjects was made in Swedish.

Details on the method will follow under separate headings.

STIMULI

The stimuli consisted of two different musical events, each recorded simultaneously with five different 5-channel microphone techniques. All recordings were reproduced through a 5-channel system, whose loudspeaker positions conformed to BS 1116 [19]. The choice of stimuli was made to follow up the discussion in a previous validation experiment [17], in which different modes of reproduction were used by the authors to create differences between stimuli. As a result of that experiment, it was suggested that a new experiment should seek to decrease the spatial differences between stimuli, e g by not altering the modes of reproduction, but instead by using different microphone techniques. In [17], the stimuli used were all single stationary centre-positioned sources within an enclosed space (a room/hall). To extend the types of sound sources in this experiment, one of the musical events used comprised two laterally displaced sound sources (a duo).

Recording techniques

In total, five different 5-channel microphone techniques were used. They were chosen to cover intensity difference and time difference principles as well as a range of different microphone directivities. The techniques are a set of earlier published as well as more informal ones. For details on microphones and their positioning, refer to figure 1.

The techniques (with their abbreviations used in this paper in italics) were:

- *card:* All spaced cardioid microphones, this particular set-up is known as the “Fukada tree” [20].

• *card8:* Frontal array: 3 spaced cardioid microphones, identical to frontal array of the *card* technique, rear array: 4 spaced bi-directional microphones, suggested by Hamasaki et al [21] and described by Theile [22].
• *coin:* Frontal array: 3 coincident cardioid microphones, rear array: 2 narrowly spaced cardioid microphones, used by the authors in [12]
• *omni:* All spaced omni-directional microphones, frontal array: microphones positioned close to the frontal array of the *card* technique, rear array: placed in the hall, away from the stage.
• *omniS:* Same as the *omni* technique, but level of each microphone in rear array raised 3 dB compared to the *omni* technique.

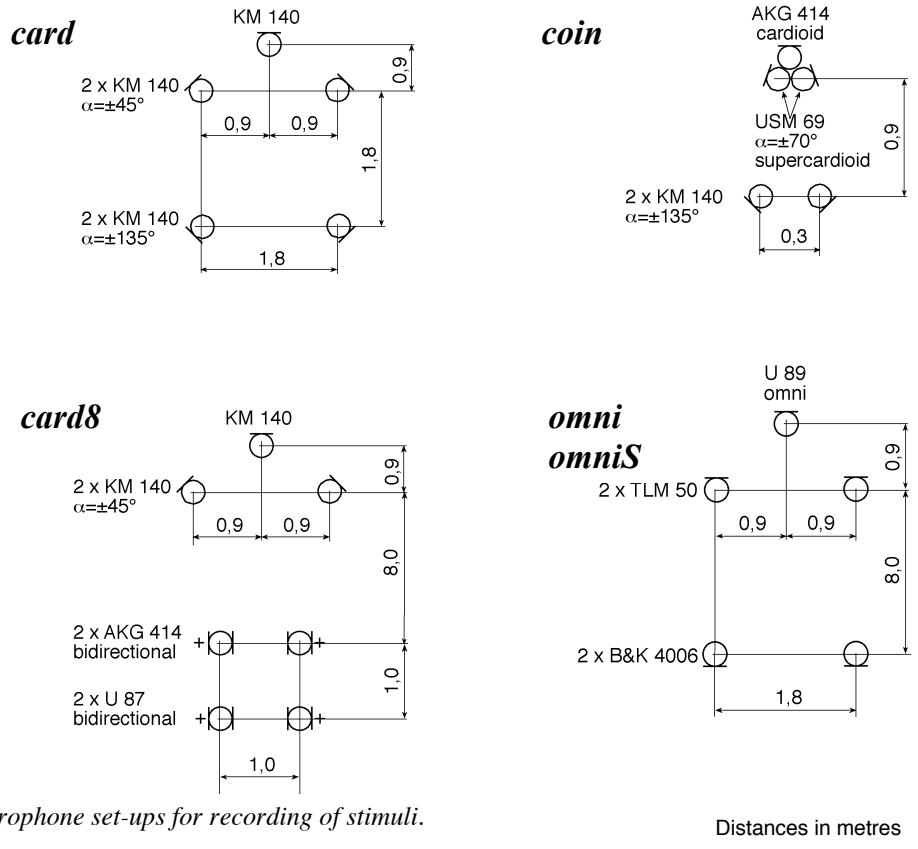


Fig. 1: Microphone set-ups for recording of stimuli.

Programmes

As mentioned above, the type of source material was expanded compared to the 2001 experiment [17], by the inclusion of both a single and a dual source as stimuli. The pieces of music are referred to as "programme" in this paper.

The programmes used (with their abbreviations used in this paper in italics) were:

- *viola:* Viola solo: G Ph. Telemann: “Fantasie für Violine ohne Bass”, e-flat, 1st movement “Dolce”. Duration: 2 minutes 19 seconds. The musician was positioned on the symmetry line of the microphone set-up, i e ‘centre-positioned’ and approximately 3 m from the closest centre microphone.
- *vocpi:* Song and piano: “Det är vackrast när det skymmer”; lyrics: Pär Lagerkvist; music: Gunnar de Frumerie. Duration: 2 minutes 18 seconds. The singer was positioned slightly right of the symmetry line of the microphone set-up and the piano slightly left of that line.

To include more than two programmes was considered, but not utilised as the resulting increase of the total extent of the experiment was regarded as being too cumbersome for the subjects.

Recording and pre-processing

Both recordings were made in the recital hall at the School of Music. The microphone signals were amplified by Yamaha HA-8 amplifiers and recorded on Tascam DA-88 machines. For editing, a ProTools system was used. The edited discrete channels were stored as *.wav-files, which later were level calibrated in the listening room. The discrete files were interlaced into 5-channel *.wav-files, one per stimulus, resulting in 10 files in total (5 recording techniques × 2 programmes).

Level calibration

To avoid level dependent differences between the stimuli, a level equalisation process was made. The primary target for this process was to minimise the level differences within a programme, i e between the different recording techniques. This was achieved by measuring the A-weighted equivalent sound pressure level, Leq(A), for the first 30 seconds of each of the five versions of a programme at the listening position, with all speakers operational, and subsequently use this measure for gain adjustment of the audio files. For minimising the level difference between programmes, two persons adjusted these ‘by ear’ to make them sound equally load. During this process, it was noted that if the inter-programme level difference was equalised using the Leq(A) method, this corresponded well with the ‘by ear’ result. Hence, the Leq(A) measure was used for all level adjustments. After level adjustment of the audio files, the measurement was repeated for confirmation that the correct gain had been applied. The maximum level difference was 1.5 dB. Results of the confirmatory measurement are to be found in figure 2. The CoolEdit software was used for the level calibration process.

Programme	Recording technique	Leq [dB(A)]
viola	card	67,4
viola	card8	67,3
viola	coin	67,5
viola	omni	67,4
viola	omniS	67,1
vocpi	card	68,0
vocpi	card8	68,1
vocpi	coin	68,6
vocpi	omni	68,1
vocpi	omniS	68,2

Fig. 2: Stimuli levels measured at listener position

SUBJECTS AND EQUIPMENT

Subjects

All subjects were students, all male, from the sound recording programme at the School of Music. All except three of them had previously participated in listening tests designed to assess the total audio quality of coding algorithms in bit-reduction systems. Six of the subjects were participants in the 2001 experiment. Apart from that, the subjects had received neither any special training in assessing spatial quality, nor any instructions in using common language for describing the spatial features of recordings. In conclusion, the subjects should be regarded as more experienced listeners of reproduced sound compared to the overall population. In the main experiment, 16 subjects participated. From this group, four subjects took part in the pre-elicitation experiment. No subject failed to complete the experiments.

Listening conditions

The experiment was executed in a reproduction room at the School of Music. The dimensions of the room was 6 □ 6.6 □ 3.2 m (w □ d □ h). All reproduction was made through Genelec 1030A loudspeakers, configured according to BS-1116 [19] at a 2 m distance from the listening position, figure 3. The settings of each loudspeaker were: Sensitivity = +6□dB, Treble tilt = +2□dB, Bass tilt = -2□dB. Only one subject at a time was present in the listening room during the experiment. Equipment with fans was acoustically insulated to avoid noise in the listening room. The room had no windows and the light in the room was dimmed. This was to increase the subject's concentration on the user interface and minimise visual distraction from the room.

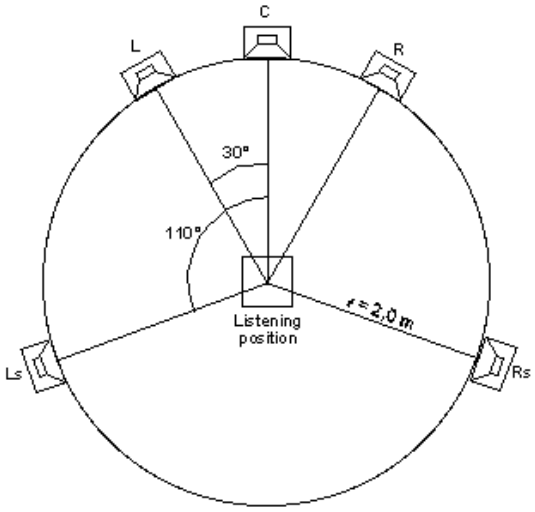


Fig. 3. Loudspeaker set-up

Reproduction equipment

The experiment was performed on a computer (PC) by which each test session was controlled. All sound files were stored on the computer's disk and played back via a Mixtreme 8-channel sound card installed in the computer. (Only five channels were used.) The sound card output delivered audio data in the T-DIF format, which was converted by a Tascam IF-88AE into the AES/EBU format, feeding a Yamaha DMC-1000 mixing console. The console was used for reproduction level adjustments and its outputs, also in the AES/EBU format, were converted by M-Audio digital-to-analogue converters to five discrete analogue signals directly feeding the speakers.

For controlling the test, special software was designed. Both playback controls as well as collecting subject responses were handled by the software. All stimuli (sound files) under test were accessible by pointing and clicking on the computer screen. The points in time between which the sound files played back were adjustable for the subject to facilitate listening between desired points and for desired durations.

ATTRIBUTES

The attributes used in this experiment waere a result of the pre-elicitation experiment (described in more detail in [18]) and the experiments previously reported by Berg and Rumsey. The purpose of the main experiment is to verify if findings about attributes elicited and tested in previous experiments still are valid under new conditions. In addition, the constructs generated in the pre-elicitation experiment are to be considered for inclusion in the main experiment. The selection of attributes for the main experiment is therefore a task of

deciding both which previously encountered attributes to keep, and which elicited within this experiment to add to the final list of attributes.

The elicitation of constructs and their refinement into attributes are described by the authors in [12, 13] (elicitation), [14] (verbal protocol analysis of subject responses) and [17] (selection of attributes and attribute list). The attributes in the 2001 experiment were divided into classes depending on whether they were describing the whole sound as an entity, the sound source (the voice/instrument only), the enclosed space in which the source was positioned (the room), or other properties. The classes were named *General*, *Source*, *Room* or *Other*. The constructs generated in the pre-elicitation experiment were now compared with the attribute list from the 2001 experiment, so that each construct was considered and subsequently associated with an attribute describing a similar property of the sound. If an association between a construct and the attributes on the list was not found, the list was augmented with a new attribute describing that construct. For some constructs, more than one attribute was associated to them, due to either the ambiguity of their meaning, or their content of more than one phrase. These interpretations were made by the author.

As the size of the main experiment is dependent on the number of attributes included, this number has to be considered carefully. An experimental design for evaluating several attributes generates many data points, with an increased risk of listener fatigue, which could result in data with low reliability. Therefore, the listener's grading consistency of the different attributes from the previous experiment, in combination with an assessment of whether certain attributes are describing spatial features of the sound or not, were used for finalising the attribute selection. Hence, the attribute list for utilisation in the main experiment consists of the following attributes with their abbreviation and their attribute class:

• *low frequency content*	*lfc*	*General*
• *naturalness*	*nat*	*General*
• *preference*	*prf*	*General*
• *presence*	*psc*	*General*
• *ensemble width*	*ewd*	*Source*
• *localisation*	*loc*	*Source*
• *source envelopment*	*sev*	*Source*
• *source width*	*swd*	*Source*
• *source distance*	*dis*	*Source*

• *room envelopment*	*rev*	*Room*
• *room size*	*rsz*	*Room*
• *room level*	*rlv*	*Room*
• *room width*	*rwd*	*Room*

Finally, as the programme *vocpi* comprised a voice and a grand piano, the subjects received additional instructions in order to focus on one of the sources at a time, when making their assessments. Given that, the source width and the localisation were each assessed twice, one time per sound source and attribute, thus resulting in the attributes *swd1*, *swd2*, *loc1* and *loc2*, where the suffix "1" indicates, in the dual source programme, that the attribute refers to the instrument (the grand piano), whereas "2" indicates a reference to the voice. The viola was assessed on all attributes. In total, 15 attributes were assessed. For description of the attributes, refer to Appendix A.

MAIN EXPERIMENT DESIGN

The framework of the main experiment was to provide a group of non-naive subjects with a list of attributes with associated descriptions and, for every attribute, let the subjects listen to sounds recorded with different recording techniques and grade the stimuli on scales defined by the attributes. The subjects performed the experiment one at a time in a listening room equipped with loudspeakers and a user interface in the form of a computer screen, a keyboard and a mouse. All communication with the subjects was made in Swedish.

Subjects

The group of subjects is described in more detail above. The number of subjects completing the main experiment was 16. No subject failed to complete the experiment.

Experimental procedure

Prior to an experiment session, every subject received a written instruction, where the experiment was described. The list of the attributes (Appendix A), to be used in the experiment accompanied the written instruction. The subjects were allowed to ask questions about the instruction, but not about the attributes and their descriptions. The instruction and the attribute list were available for the subjects during the whole session.

A session started with a training phase where only four of the attributes were included to avoid subject fatigue at the end of the test. The purpose of the training phase was to familiarise the subjects with the equipment and the stimuli used in the test.

Each subject was first presented a computer screen with text showing one attribute with its description. In addition to that, all 10 stimuli (two programmes recorded with five recording techniques) were available for listening by clicking on buttons on the computer screen. The task was to grade all stimuli one by one on the attribute presented. This was accomplished by providing 10 upright continuous sliders on the screen, one slider per stimulus. The subjects were instructed to regard the scale on the sliders as linear. The slider had two markings only, one at each endpoint, the lower marked "0" (zero) and the upper marked "MAX". The subject was also instructed to use the MAX grade for at least one stimulus, but did not necessarily have to give any stimulus the grade 0. When the subject was satisfied with his grading on the first attribute, the scores were stored by clicking a button, whereupon the next attribute was presented. All stimuli were graded again, but now on the new attribute. This was repeated until all attributes were graded by the subject. When this was completed, the session finished.

To avoid systematic errors, the presentation order and assignation of playback buttons were randomised: When a session started, the attribute class was chosen randomly. The order in which the attributes within the chosen class were presented was also picked randomly. When all attributes within the class were assessed by the subject, a new attribute class out of the remaining ones was randomly chosen. This was repeated until all attribute classes with their attributes were assessed. For every new attribute, the assignation of the stimuli to the 10 playback buttons was re-randomised. In total 15 trials, one per attribute, were made per session and subject.

Data acquisition

The slider position representing a subject's assessment of a given stimulus on a given attribute was converted into an integer number from 0 to 100, where 0 corresponds to the marking "0" and 100 to "MAX", and the intermediate values are equally distributed on the length of the slider. The converted grades with proper identification of subject, associated stimulus, attribute and date/time were stored on the computer in one text file per subject. The text files were later converted into MS Excel files for subsequent loading into the statistical analysis software.

INTRODUCTORY DATA ANALYSIS

Before commencing the different planned analyses, the experimental data is subjected to transformation and testing for basic statistical properties.

Data structure

The data acquired consisted of 16 subjects assessing 10 stimuli on 15 attributes, yielding 2400 data points. Every subject delivered 150 grades.

Data transformation

As the scale used for the grades is not absolute and does not contain any absolute anchors (apart from "0"), in order to facilitate the comparison of grades between stimuli across subjects, the subjects' different use of the scales provided must be equalised. This is accomplished by, for each subject, normalising the grades given to an attribute. This way, the grades given to each attribute are transformed to have the same mean value and the same standard deviation as the other attributes for all subjects. The operation also removes the subject (listener) effect from the following analyses. There are 10 stimuli per attribute and the mean value

$$\bar{x}_{ik} = \frac{1}{10}\sum_{j=1}^{10} x_{ijk}$$

and the standard deviation

$$s_{ik} = \sqrt{\frac{1}{9}\sum_{j=1}^{10}(x_{ijk} - \bar{x}_{ik})^2}$$

where

x_{ijk} = grade given on attribute i for item j by subject k

are used for calculating the z-score

$$z_{ijk} = \frac{x_{ijk} - \bar{x}_{ik}}{s_{ik}}$$

which now is the normalised value of the original grade. The mean value of z-scores per subject and per attribute is 0 and the standard deviation is 1. Consequently, the data now consists of normalised values in the form of z-scores suitable for the coming steps in the analysis.

Data properties

To examine if the z-scores given for each stimulus on each attribute are normally distributed across subjects, Shapiro-Wilk's test [23] is performed. Since 16 subjects graded 10 stimuli on 15 attributes, the number of cases to be tested is 150, each containing 16 observations. The outcome of this test, expressed as probabilities for normal distribution for the different cases, when the level of confidence is set to 95%, the test shows that a normal distribution cannot be excluded in 125 of the 150 cases. The observations seem to be normally distributed in more than 80% of the cases, which indicates some consistency between the subjects in their grading of the stimuli. Normal distribution also an assumption underlying Analysis of variance (Anova).

Another assumption underlying Anova is the homogeneity of the variances of the data in each cell (5 recording techniques × 2 programmes = 10 stimuli = 10 cells). Thus, for every attribute, there are 10 cells, which variances of the z-scores are compared by Cochran's C test. At a confidence level of 95%, all attributes except the *ensemble width*, *ewd*, pass the test. This means that, in this respect, Anova can be used for finding significant differences among the mean values, except for the *ewd* attribute. However, Lindman [24] shows that the *F* statistic is quite robust against violations of this assumption and therefore *ewd* is also subjected to Anova.

ATTRIBUTES' DISCRIMINATION POTENTIAL

There are two main purposes of the analysis. Firstly, to establish if the provided attributes enable the group of subjects to significantly discriminate between different recording techniques. Secondly, if discrimination between the recording techniques is found, to determine which techniques are significantly separable by the different attributes. Of interest are also how consistent the group of subjects is in its assessment of the different attributes, and if the type of musical event is a significant factor in the analysis. Since normal distribution and equal variances were not excluded by the introductory analysis apart from in a few cases, Analysis of variance is used for finding differences between the mean values of

the cases of interest. A factor is considered significant when its *F*-ratio has a probability $p<0.05$.

Significance of attributes

The significance of each attribute is tested by means of Analysis of variance (Anova) of the z-scores given to the stimuli. In the Anova model, the dependent variable is the normalised grade (*z-score*) and the factors are recording technique (*rec_tech*) and the type of musical event (*programme*). The interaction between the two factors is also included in the model. The factor *rec_tech* comprises five levels and the factor *programme* two levels. Since the data was normalised as described above, the *F*-ratio of the factor subject (*subid*) is zero, which confirms that the subject effect is removed from the analysis, as intended. For each attribute and factor, the definition of the null hypothesis

H_0 : No significant difference is found between the mean values of the factor levels, which indicates that the attribute provided is not sufficient for enable the subjects to find a significant difference between any stimuli

and the alternative hypothesis

H_A : A significant difference is found between the mean values of the factor levels, which indicates that the attribute provided is sufficient for enable the subjects to find a significant difference between at least one stimulus and the other stimuli

For the main effect of the factor *rec_tech*, the analysis shows that for all 15 attributes, the *F*-ratios correspond to significance levels $p<0.001$, except in one case, the attribute *presence*, where $p<0.05$. The null hypothesis is therefore rejected for *rec_tech*, in favour of the alternative hypothesis for every attribute. Hence, for all attributes, there are mean values of grades given to recording techniques significantly differentiating, thereby showing the attributes sufficient for making distinctions between some recording techniques. The attributes must therefore have some common meaning to the subjects; otherwise, the individual subject differences would have resulted in randomly distributed data points across the stimuli, yielding insignificant differences in means between the stimuli. The Anova tables are found in Appendix B.

The main effect of the factor *programme* is significant ($p<0.05$) for 7 of the 15 attributes. These are (with their abbreviation and attribute class):

- *low frequency content* *lfc* *General*
- *preference* *prf* *General*
- *ensemble width* *ewd* *Source*
- *localisation1* *loc1* *Source*
- *source envelopment* *sev* *Source*
- *source width1* *swd1* *Source*
- *source distance* *dis* *Source*

For the remaining 8 attributes, the main effect of the factor *programme* is not significant:

- *naturalness* *nat* *General*
- *presence* *psc* *General*
- *localisation2* *loc2* *Source*
- *source width2* *swd2* *Source*
- *room envelopment* *rev* *Room*
- *room size* *rsz* *Room*
- *room level* *rlv* *Room*
- *room width* *rwd* *Room*

For the attributes showing non-significant *F*-ratios of the factor *programme*, the interaction between *rec_tech* and *programme* is examined for which combinations of them significant interactions occur. This is accomplished by a follow-up test, comparing mean values of programmes on each recording technique and searching for differences, exceeding the Tukey Honestly Significant Difference (HSD) interval (which is chosen for reducing the risk of Type I errors when performing multiple comparisons, as described in [25]). Only for *presence* and *room envelopment* such a difference is found for the *card8* recording technique. The rest of the attributes having non-significant *F*-ratios for *programme* do not show any programme dependent differences between recording techniques exceeding the Tukey HSD.

Examining the main effect of the factor *programme*, in most of the *Source* attribute class, it is a significant factor, whereas for all four attributes in the *Room* attribute class, it is not. The latter seems to support that the characteristics of the room in most cases can be perceived and assessed regardless of the type of source (apart from *rec_tech* = *card8*). Neither *naturalness* nor *presence* are attributes for which *programme* is significant factor (apart from

rec_tech = card8 for *presence*, as noted above). This could be because both sources are naturally existing musical events, both giving the same sensation of presence in most cases.

The two *Source* attributes with the suffix 2 refers in the dual source case (song and piano) to the voice, i e the 'narrower' of the two. The result indicates that the voice is perceived more similar to the other programme, the solo viola, in terms of width and localisation, and therefore cannot be separated by *loc2* and *swd2*. However, for *loc1* and *swd1*, referring to the piano in the dual source case, *programme* is a significant factor, which shows that the piano is perceived as having another width and localisation than the viola.

The *F*-ratio for interaction between the factors is significant for all attributes, with the exception of *naturalness*. This indicates that there are certain combinations of recording techniques and programmes that are perceived significantly different from other combinations of the two factors on the same attribute. Graphs depicting the interactions are found in Appendix C and a summary of these showing the attributes able to bring out differences between recording techniques within each programme is in figure 4. From this is noted that the programme *vocpi* enables the group of subjects to discriminate between recording techniques on all attributes, whereas *viola* does so for 9 of the 15 attributes. However, since the recording techniques in themselves show to be significantly different, this is sufficient for rejecting the null hypothesis for the factor *rec_tech*, thereby concluding that the group of subjects can discriminate between certain recording techniques for all attributes. Which of the recording techniques this applies to is analysed in the follow-up test in the following section.

Attribute	**Significant difference between *rec_tech* within *programme***	
	viola	**vocpi**
lfc		*
nat		*
prf	*	*
psc		*
dis	*	*
ewd	*	*
loc1		*
loc2	*	*
sev		*
swd1	*	*
swd2		*
rev	*	*
rlv	*	*
rsz	*	*
rwd	*	*

Fig. 4: Significant differences between recording techniques for each programme and attribute. Tukey's HSD is used for all attributes, except ewd*, where 95% confidence intervals calculated from individual standard errors are used.*

Comparison of recording techniques

As the factor *rec_tech* is found to be significant, the mean values of the *z*-scores given to different recording techniques can be compared to find the means significantly different. For all attributes passing the equal variance test (14 out of 15), the multiple range tests with Tukey HSD intervals ($p<0.05$) is used [25],while the remaining attribute (*ensemble width*) is subjected to comparison of mean values for recording techniques with their associated individual 95% confidence intervals, derived from their individual standard errors. However, interpretations of means must be made carefully, as significant interactions with *programme* were found in the foregoing section. Graphs showing the interactions are in Appendix C. When making the following comparisons of the main effect of *rec_tech*, some remarks on the attributes can be made: *coin – omniS* are separable by all attributes; *omni – omniS* are separable only by *room width*, and *card – card8* are separable only by *room width* and *localisation2*. The attribute *presence* is able only to bring out a difference for *coin – omniS*, but not for any other comparisons between techniques. No attributes in the *Source* class are sensitive to the *omni – omniS* difference (which is a 3 dB change of the rear speakers level). If *localisation2* is disregarded, this lack of sensitivity for *Source* attributes applies to *card – card8* too. Common for these two comparisons are that the frontal microphone array is identical within each comparison. In the *card8* technique, two of the rear array microphones are mixed into the signals feeding the front left and right speakers, evidently causing a difference detectable by the attribute *localisation2*, which represents the ability to localise the narrow sources (voice and viola). A study of the number of differences between all possible combinations of stimuli, i e taking the interaction of recording techniques and programmes into account, shows that in 6 out of 45 comparisons there is no significant difference between stimuli. This applies to the following pairs: card(viola) – omni(viola); card(viola) – omniS(viola); card8(viola) – omniS(viola); coin(viola) – coin(vocpi); omni(viola) – omniS(viola) and card(vocpi) – card8(vocpi). A low number of differences are also predominant for other comparisons within the stimuli comprising the viola. Evidently, the attributes used are less sensitive to differences between techniques for this type of programme.

Consistency in attribute grading

To evaluate the quality of an attribute as a mean of both describing a certain feature of the sound as well as creating a common interpretation of the feature, the consistency in grading within the group of subjects is analysed for each attribute. A relatively high consistency is likely to indicate a more similar perception of the attribute than a relatively low one. To test this, the residual (or error) variance for each attribute are taken from the Anova and compared to the other attributes' residual variances. Since the between-subject variability was removed earlier from the Anova model by the normalisation procedure, the residual variance only consists of the differences in magnitude and direction of the trends in subject performance. Consequently, a low residual variance indicates a high consistency in trends [25]. The residual variances are shown in figure 5.

When the attributes' residual variances are ordered in ascending order and these variances are inspected, the most consistently graded attributes are *source width1* and *low frequency content*, whereas the least consistently graded are *naturalness* and *presence*.

Some observations on these results, when compared with those from the 2001 experiment [17], are made. *Naturalness* shows low consistency in both experiments, indicating larger differences in individual appreciation of this attribute. *Preference* changes from high to low consistency, which presumably is a result of that, in the 2001 experiment, a number of mono reproductions were used as stimuli, which differed more noticeably from the non-mono stimuli, resulting in more consistent preferences for the latter.

Attribute	Residual variance
swd1	0,36671
lfc	0,36867
sev	0,41760
rlv	0,51530
ewd	0,51881
dis	0,53885
loc1	0,56345
rwd	0,59344
rsz	0,60386
rev	0,61558
prf	0,61944
swd2	0,70524
loc2	0,71122
psc	0,77390
nat	0,80647

Fig. 5: Residual (error) variances for attributes

CORRELATION AND DIMENSIONALITY OF ATTRIBUTES

An important part of evaluating the attributes is to examine their interrelation. If attributes are scored similarly on the different stimuli, it is an indication of that they are perceived in a similar way. On the other hand, if there are attributes showing to be independent, this is an important finding for understanding the dimensionality of the data generated by the subjects' perception of the stimulus set. For exploring the interrelations, correlation analysis and factor analysis are performed on the data.

Correlation analysis

To find the correlation in terms of the linear relationship between the attributes, the Pearson product moment correlation coefficient, r was calculated [26]. The results are given as a coefficient for every pairwise combination of the attributes. The correlation coefficients and their p-values are found in Appendix D. If $r = 0$ for a pair of attributes, no linear relationship exists between these [27]. When $r \neq 0$, a correlation exists if the difference from zero is significant. The interpretation of the coefficients is based on the informal definition by Devore and Peck [26], where the magnitude of r is considered as an indicator of the strength of the linear relationship as follows: $|r| \leq 0.5$ is a weak, $0.5 < |r| \leq 0.8$ is a moderate and $|r| > 0.8$ is a strong relationship. Using this terminology, a number of observations are made.

No strong relationships are found. In six cases moderate relationships are found. Significant correlations ($p \geq 0.05$) do not exist in 26 of the comparisons. The rest of the comparisons show significant but low correlations. The moderate relationships are found between these attributes:

- *source envelopment – low frequency content*
- *source width1 – low frequency content*
- *source width1 – ensemble width*
- *source width1 – localisation1 (negative)*
- *source width1 – source envelopment*
- *source distance – room level*

Obviously, the group of subjects consider the properties described by the *source width1* attribute similar to other width attributes, like the envelopment of the source (the piano) and the width of the ensemble. As the source is perceived to get wider, the ability to localise the source drops, as encountered in the Berg and Rumsey's previous work [17], where *source*

width and *localisation* have a correlation coefficient of –0.602. A similar relation has also been confirmed recently by Zacharov and Koivuniemi [28], where their attributes *broadness* and *sense of direction* show a correlation of –0.587. The remaining moderate relationship indicates that a greater distance to the source seems to coincide with a higher level of the room sound, which presumably is a detection of the direct-to-reverberant sound ratio.

The attributes showing the highest number of uncorrelated other attributes are *source distance* and *localisation2*. Each of them lacks a significant correlation to eight other attributes. The correlation between *source distance* and *localisation2* are negatively weak (r=-0.33).

Looking at the attributes within each attribute class, the attributes within the *General* class show to be significantly but low correlated. This applies to the attributes in the *Room* class too. Hence, the attributes within each of these classes are not completely independent. Most of the *Source* class attributes are non-correlated with some other attributes within the *Source* class. This is salient for *source distance* and *localisation2*, which each lacks correlation with three other attributes, all describing forms of width, within the *Source* class.

For exploring if a pattern of the remaining uncorrelated attributes can be discovered, the correlations between attributes belonging to different attribute classes are studied for the lack of significant correlation. When inspecting correlations between attributes in the *General* and the *Source* classes, 10 uncorrelated pairs of attributes are found. All of them comprise localisation and distance attributes, which implies that these do not form the basis on which the more general (or holistic) attributes are perceptually derived. Repeating this procedure for the attributes in the *General* and the *Room* attribute classes shows that *room level* is uncorrelated with three of the four general attributes. It is noted that these three attributes in the *General* class (*naturalness, preference* and *presence)* all can be characterised as being attitudinal rather than descriptive, as discussed in previous work [14]. Finally, inspecting non-correlation between attributes in the *Room* and the *Source* attribute class reveals that *room envelopment* is uncorrelated to the *source distance* and both *localisation1/2* attributes. The attribute *ensemble width* is uncorrelated to both *room level* and *room size*. For *source distance* and *room with*, there is no correlation.

Factor analysis – all attributes

Factor analysis (FA) is used when an accurate description of the domain covered by the variables is desired. This is chosen in favour of principal component analysis (PCA), since

the extraction of components in a PCA considers all variance, so the components are likely to consist of more complex functions of the variables (than a FA), which could make the components harder to interpret [2]. The factor analysis is performed on the set of attributes, which corresponds to the columns in the matrix of the z-scores

$$\mathbf{A} = \begin{bmatrix} z_{1,1,1} & \cdots & z_{15,1,1} \\ \vdots & & \vdots \\ z_{1,jk} & \cdots & z_{15,jk} \\ \vdots & & \vdots \\ z_{1,10,16} & \cdots & z_{15,10,16} \end{bmatrix}$$

where

z_{ijk} = z - score on attribute i for item j by subject k

and where the matrix's columns were normalised prior to the FA. The number of factors is determined by the Kasier criterion, which states that all components with an eigenvalue [illegible] should be kept in the analysis. Applying this, three factors are extracted in the analysis, accounting for 58 % of the variance. The eigenvalues and variances are shown in figure 6.

Factor Number	Eigenvalue	Percent of Variance	Cumulative Percentage
1	5,09284	33,952	33,952
2	2,14921	14,328	48,280
3	1,42081	9,472	57,752
4	0,89306	5,954	63,706
5	0,76994	5,133	68,839
6	0,73652	4,910	73,749
7	0,69275	4,618	78,368
8	0,60335	4,022	82,390
9	0,52942	3,529	85,919
10	0,50366	3,358	89,277
11	0,42092	2,806	92,083
12	0,39956	2,664	94,747
13	0,32309	2,154	96,901
14	0,26261	1,751	98,652
15	0,20226	1,348	100,000

Fig.6: Eigenvalues and cumulative variances of the factors

To increase the interpretability, the factors are rotated, using Varimax, to maximise the loadings of some of the attributes. These attributes can then be used to identify the meaning of the factors [30]. The loadings on the extracted factors are presented in figure 7.

Attribute	Factor 1	Factor 2	Factor 3
lfc	0,7162	0,3467	0,1042
nat	0,0729	0,6645	0,0244
prf	0,3012	0,6873	-0,2589
psc	0,1109	0,6325	0,0228
dis	0,0726	-0,1489	0,8222
ewd	0,7475	0,1877	-0,0763
loc1	-0,6632	0,1467	-0,4390
loc2	-0,0777	0,0186	-0,6018
sev	0,7547	0,2977	0,0246
swd1	0,8407	0,2104	0,1967
swd2	0,4263	0,4569	0,1802
rev	0,2475	0,7013	0,1320
rlv	0,1400	0,2125	0,7646
rsz	0,0153	0,4266	0,6130
rwd	0,3552	0,5562	0,3430

Fig. 7: Loadings on the three extracted factors by the attributes

To understand the factors in terms of the attributes, the procedure described by Bryman and Cramer [30] is utilised. The procedure is distinguished by, for each factor, selecting the variables (the attributes) having a loading greater than 0.3 on that factor uniquely, as the variables characterising the factor. Applying this, the following is observed about the factors.

- Factor 1 is characterised by *ensemble width*, *source envelopment* and *source width1*. This is clearly a width factor referring to the source primarily. If the constraint of unique loading on one factor is dropped, *location1* is included and loads factor 1 negatively.
- Factor 2 is characterised by *naturalness*, *presence* and *room envelopment*. This factor seems to account for the sense of being present at the venue where the sound source is, and at the same time, it also seems to indicate that it is the enveloping room that forms a part of this conception. Dropping the unique loading constraint, the other attributes in the *Room* class, except *room level* also become included and load this factor too.
- Factor 3 is characterised by *room level* and *source distance*, and on the negative part, by *location2*. Considering the attributes on the factor, this is a general distance factor; as the source distance increases, the room level does. At its negative end, the existence of *localisation2* could imply that when the distance decreases, the source is easier to localise, perhaps due to a lower level of reverberation. The attribute *room size* loads this factor as well as factor 2. A speculation, since no width attributes load this factor strongly, is that this is a factor representing a conception that 'works' in mono too.

Plots showing the loadings on the factors are in Appendix E.

To find the way in which the techniques used for recording the programmes relate to the extracted factors, the factor scores are examined. For each factor, the highest (most positive)

25% and the lowest 25% (most negative) of the factor scores are filtered out and each of these factor scores is analysed for which recording technique it represents. (25% equals 40 factor scores.) The number of occurrences of different recording techniques is counted for each factor. Since both high (positive) and low (negative) factor scores are selected and analysed, both endpoints of each factor thereby are associated with the recording techniques most applicable for the factor. The number of occurrences for each technique is the table in figure 8 and from this, the following is noted:

- Both factor 1 and factor 2 show the most positive factor scores for the both omnidirectional techniques (*omni* and *omniS*) and the most negative factor scores for the coincidence technique (*coin*).
- The scores on factor 3 are most positive for the cardioid techniques (*card* and *card8*) and most negative for the coincident technique (*coin*).

Rec_tech	F1 H	F1 L	F2 H	F2 L	F3 H	F3 L
card	2	4	1	6	10	2
card8	3	1	8	7	22	1
coin	0	28	0	23	0	27
omni	16	4	12	2	1	7
omni8	19	3	19	2	7	3

Fig. 8: Distribution of the highest (H) 25% and the lowest (L) 25% of the factor scores on each factor (F). Table shows number of factor scores associated with the different recording techniques

Conclusions on the recording techniques

Combining the results of the factor loadings and the factor scores, the following can be concluded. The omni-directional techniques create a sound characterised by a greater width and a poorer localisation of the source. Good detection of presence and prominent reverberation envelopment are also typical of these techniques. The coincidence technique has a low amount of these features, whereas it gives a good localisation of the sources and closeness to them. The cardioid techniques, especially the *card8*, result in a distant and reverberant sound.

At this stage, it has to be emphasised that the experiment's purpose was to detect differences between recordings resulting from a variety of recording techniques, but not to make absolute and generalisable judgements on a particular recording technique.

Factor analysis – emphasis on room attributes

The notion of being present at the scene of the auditory event and the characterisation of sounds as "natural", correlates weakly with some, but not all, of the attributes describing the room/hall. There are also weak, but still significant, correlations between the attributes in the *Room* class. This is apparent, both in this and in the 2001 experiment [17], and the question of what constitutes "presence" in a reproduced sound emerges: Which of the room attributes contributes to presence and which are most likely independent from this? To get a clearer picture, the attributes in question were examined by means of factor analysis. The analysis was made on the four attributes in the *Room* class: *room envelopment*, *room level*, *room width* and *room size* plus the attribute *presence*. This was achieved by including only the columns of the matrix **A** containing these attributes. Two factors were extracted, as a result of employing Kaiser's criterion. Varimax rotation was used also in this analysis. The plot of the factor loadings is in figure 9.

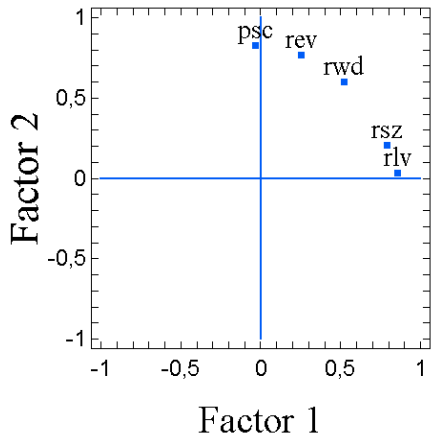


Fig. 9: Factor loadings of room attributes only. Two factors extracted. Rotation: Varimax.

The plot of the factor loadings suggests, on the first factor, that *room size* and *room level* are attributes describing one underlying dimension, whereas on the second factor, *presence* and *room envelopment* are describing another. The remaining *room width* describes a combination of these two dimensions. The author of this paper is proposing that the perception of the enclosed space can be divided into a judgement dimension and a sensation/impression dimension. A perception within the judgement dimension is characterised by the ability to judge or determine some properties of the environment, the room, the hall or the enclosed space in which the sound source is positioned. Examples of

this are the size of the space and the level of the reverberation. The sensation/impression dimension is represented by a sense of actually being present in the acoustical environment, within the room/hall/space. The difference between these dimensions is that attributes in the judgement dimension do not require an impression of presence to be perceived and determined.

To see if similar results could be observed in data from other experiment(s), findings in the present experiment were compared with an, until now, unpublished analysis of data from the 2001 experiment [17]. The same type of factor analysis described in this section is utilised on the 2001 experiment's data, associated with the same attributes, figure 10. The analysis shows that a similar pattern exists in both experiments. (In the previous experiment, the attribute *envelopment*, *env*, was not separated into two separate attributes referring either to the source or to the room. It was instead considered as a general attribute.)

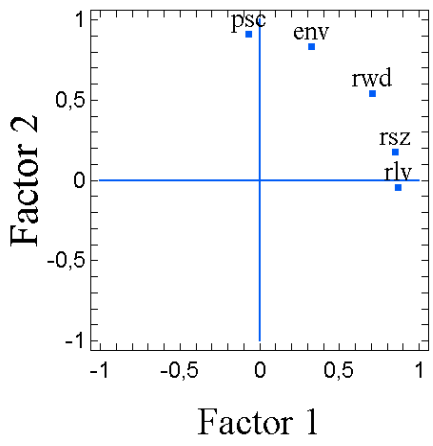


Fig. 10: Factor loadings of room attributes only in the 2001 experiment. Two factors extracted. Rotation: Varimax.

CONCLUSIONS AND DISCUSSION

Summary of results

The results, given the conditions in this experiment, can be summarised as follows:

- Subjects are able to find spatial differences between different recording techniques by comparing them in triads.
- Comparison of reproduced sound stimuli utilising triads can be used for eliciting constructs in the form of verbal descriptors.
- Grading of previously elicited attributes of reproduced sound accompanied by descriptions in writing can be used for finding spatial differences between different recording

techniques. This enables an assessment of the stimuli on the properties described by the attributes.

- When assessing stimuli, the group of subjects can focus on different aspects of reproduced sound, such as perceived properties of either the sound source or the space that the source interacts with.
- Attributes referring to the space (the room) seem to be judged independently of the type of sound source in most cases.
- The attributes used seem to be less sensitive to differences between the stimuli comprising the viola.
- No strong linear relationships are found between the attributes.
- Some attributes show a non-significant correlation with other attributes. This is predominant for the source distance and localisation attributes.
- The attributes used seem to be perceived mainly in three dimensions; width, distance to the source and a sensation of presence in the room/hall.
- The attributes describing the space/room are perceived in a judgement dimension and a sensation/impression dimension.
- Some observations on the different recording techniques perceived features are made, e g the omnidirectional techniques emphasise width, whereas the coincident technique gives better localisation of the sound source. (However this experiment was not primarily concerned with a comparison between recording techniques, and the techniques concerned have not necessarily been compared under the most suitable or favourable conditions in each case.)

Discussion

As the aim of the work in this paper concerns understanding of subjective features constituting spatial quality, it has to be noted that the classification of attributes as spatial or non-spatial is a matter of definition. The elicitation method used does not in itself exclude any constructs, unless constraints are put on parts of the elicitation process. Examples of constructs that could be regarded as non-spatial are constructs referring to the frequency spectrum or different types of attitudinal constructs. Somewhere in the process of finding certain types of attributes, a decision on the classification of these has to be made by someone. This decision process obviously influences the final result. Some of the issues regarding the interpretation of verbal data are discussed in a previous paper [14]. In this

experiment, in the process of deciding which attributes should be included in the main experiment, the interpretation of the relation between the elicited constructs and the existing attributes was made by one person. To decrease the bias risk in future applications of this method, this stage could be performed by a group of people, thus averaging out extreme differences in interpretation.

As noted already in the 1998 experiment, subjects indicate that certain stimuli give them a feeling of presence in the space (the room/hall) where the sound source is. This feeling appears to be more related to attributes referring to the space than to the sound source. The results from the experiment reported in this paper, as well as the results in the 2001 experiment, suggest that the perception of room attributes and the feeling of presence are divided into a judgement and a sensation/impression dimension. In the factor analysis, the envelopment of the listener by the room sound (e g reverberation) is within the same dimension as the feeling of presence, which implies that this form of envelopment is important for the experience of presence.

This is the second experiment where a group of subjects use attributes originating from individually elicited constructs to evaluate a set of stimuli. The results show that listeners with an above average experience of listening to reproduced sound can use selected verbal attributes defined in writing for making judgements about different recording techniques. Also the pre-elicitation experiment preceding the main experiment in this paper offers results from which conclusions about the similarities and the differences between the stimuli can be made. It is notable that all the selected attributes gave rise to statistically significant differences between stimuli, a fact that is considered unlikely had the attributes not been based on constructs elicited specifically for such spatial audio stimuli. In other words, the elicitation of *relevant* constructs for subjective evaluation is an important precursor to the evaluation itself, in order to avoid the possibility that one's chosen constructs might otherwise be of only limited relevance to the stimuli in question.

The use of attributes for evaluation of different aspects of reproduced sound is not a novel concept. It has been proposed by Bech [31] and in different standards such as IEC 60268 [32] and EBU 562-3 [33]. Experiments where attributes are used for evaluation are published by Gabrielsson and Sjögren [34], Toole [35] and Martin et al [36]. The results in the 2001 experiment [17] as well as in the present experiment, both wherein attributes successfully were used for assessment of stimuli, confirms that attributes are meaningful as tools of

focusing listeners' attention towards perceivable properties of reproduced sound, also in the case of evaluation of spatial quality.

The difference from most of the work done by others is the method used in the series of experiments (reported by Berg and Rumsey in [12, 13, 14, 15, 17]), which employ the stimuli under test for eliciting information subsequently structured and used for defining the scales upon which the stimuli are rated. A similar approach, but with a different elicitation method is used by Zacharov and Koivuniemi [28].

The conclusion, under the conditions of the experiments, is that the attributes developed as a result of an elicitation process aided by the stimuli under test are valid in the context of evaluation of stimuli differing in modes of reproduction as well as in recording techniques.

Further work

The refinement of attributes can be taken further, either by employing alternative elicitation methods or developing more precise descriptions of existing attributes to reduce the risk of overlap between them. As suggested in a previous paper, the creation of reference stimuli is also a possible way of making the meaning of the attributes more precise.

To examine the applicability of existing or recently derived attributes, the stimulus set can be altered. Besides different modes of reproduction and different recording techniques, stimuli can also differ in other ways. The programme set can be extended to comprise a higher number of sources than those occurring in the single and the dual cases used in this experiment. Another option, possibly generating smaller differences between stimuli, is to keep all factors (e g mode of reproduction, recording technique, programme) constant and assess different loudspeaker types, either by their working principle or within the same principle, different brands. Furthermore, different post-production equipment, such as reverberation systems or spatial enhancers in general, can be evaluated.

A field not yet looked into by the authors, is where some quantifiable physical parameter of the stimuli is varied while subjects' responses on scales defined by the extracted attributes are recorded. The work so far has been primarily concerned with the structuring and analysis of subjective data.

ACKNOWLEDGEMENTS

The author wish to thank Dr Francis Rumsey for his invaluable support during the series of experiments leading to the results reported here. Also the students at the School of Music,

Piteå, Sweden are thanked for their participation in this experiment, both as subjects and as musical performers. Jonas Ekeroot, JEK Sound Solutions, is thanked for his diligent programming of the software controlling the test equipment. The author also wishes to thank Andreas Renhorn, currently a student in Audio Engineering, for performing the recordings of the stimuli used in this paper. A part of the work preceding this paper was performed within the Eureka project 1653 Medusa (Multichannel enhancement of Domestic User Stereo Applications). The members of this project are thanked for their comments and discussion.

REFERENCES

1 Rumsey, F. (1998) Subjective assessment of the spatial attributes of reproduced sound. In *Proceedings of the AES 15th International Conference on Audio, Acoustics and Small Spaces*, 31 Oct–2 Nov, pp. 122–135. Audio Engineering Society

2 Kelly, G. (1955) *The Psychology of Personal Constructs*. Norton, New York.

3 Danielsson, M. (1991) *Repertory Grid Technique*. Research report. Luleå University of Technology. 1991:23

4 Fransella, F. and Bannister, D. (1977) *A manual for Repertory Grid Technique*. Academic Press, London.

5 Stewart, V. and Stewart, A. (1981*) Business applications of repertory grid*. McGraw-Hill, London.

6 Stone, H. *et al* (1974) Sensory evaluation by quantitative descriptive analysis, *Food Technology*, November, pp 24-34.

7 Mattila, V. V. (2001) Descriptive analysis of speech quality in mobile communications: Descriptive language development and external preference mapping. Presented at *AES 111th Convention, New York*. Preprint 5455.

8 Koivuniemi, K., Zacharov, N. (2001) Unravelling the perception of spatial sound reproduction: Language development, verbal protocol analysis and listener training. Presented at *AES 111th Convention, New York*. Preprint 5424.

9 Wenzel, E. M. (1999) Effect of increasing system latency on localization of virtual sounds. In *Proceedings of the AES 16th International Conference on Spatial Sound Reproduction, 10–12 Apr*. Audio Engineering Society. pp 42-50.

10 Mason, R., Ford, N., Rumsey, F. and de Bruyn, B. (2000) Verbal and non-verbal elicitation techniques in the subjective assessment of spatial Sound Reproduction. Presented at *AES 109th Convention, Los Angeles*. Preprint 5225.

11 Ford, N., Rumsey, F., de Bryun, B. (2001) Graphical elicitation techniques for subjective assessment of the spatial attributes of loudspeaker reproduction – a pilot investigation. Presented at *AES 110th Convention, Amsterdam*. Preprint 5388.

12 Berg, J. and Rumsey, F. (1999) Spatial attribute identification and scaling by Repertory Grid Technique and other methods. In *Proceedings of the AES 16th International Conference on Spatial Sound Reproduction, 10–12 Apr*. pp 51-66. Audio Engineering Society.

13 Berg, J. and Rumsey, F. (1999) Identification of perceived spatial attributes of recordings by repertory grid technique and other methods. Presented at *AES 106th Convention, Munich*. Preprint 4924.

14 Berg, J. and Rumsey, F. (2000) In search of the spatial dimensions of reproduced sound: Verbal Protocol Analysis and Cluster Analysis of scaled verbal descriptors. Presented at *AES 108th Convention, Paris*. Preprint 5139.

15 Berg, J. and Rumsey, F. (2000) Correlation between emotive, descriptive and naturalness attributes in subjective data relating to spatial sound reproduction. Presented at *AES 109th Convention, Los Angeles*. Preprint 5206.

16 Zacharov, N. and Koivuniemi, K. (2001) Unravelling the perception of spatial sound reproduction: Techniques and experimental design. In *Proceedings of the AES 19th International Conference on Surround Sound, 21-24 Jun.* pp 272-286. Audio Engineering Society.

17 Berg, J. and Rumsey, F. (2001) Verification and correlation of attributes used for describing the spatial quality of reproduced sound. In *Proceedings of the AES 19th International Conference on Surround Sound, 21-24 Jun.* pp 233-251. Audio Engineering Society.

18 Berg, J. and Rumsey, F. (2002) Validity of selected spatial attributes in the evaluation of 5-channel microphone techniques. *AES 112th Convention, Munich*. Preprint 5593.

19 ITU-R (1996) *Recommendation BS.-1116, Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems*. International Telecommunication Union.

20 Sawaguchi, M. and Fukada, A. (1999) Multichannel sound mixing practice for broadcasting. In *Proceedings of the IBC Conference 1999*. IBC

21 Hamasaki, K., Fukada, A., Kamekawa, T. and Umeda, Y.(2000) A concept of multichannel sound production at NHK. In *Proceedings of the 21st Tonmeistertagung 2000*. VDT

22 Theile, G. (2001) Natural 5.1 music recording based on psychoacoustic principles. In *Proceedings of the AES 19th International Conference on Surround Sound, 21-24 Jun.* pp 201-229. Audio Engineering Society.

23 Nelson, P. R. (1990) Design and analysis of experiments. In *Handbook of statistical methods for engineers and scientists*. Editor: Wadsworth, H. M. McGraw-Hill.

24 Lindman, H. R. (1974) *Analysis of variance in complex experimental designs*. Freeman, San Fransisco.

25 Roberts, M. J. and Russo, R. (1999) *A student's guide to analysis of variance*. Routledge, London.

26 Devore, J. L. and Peck, R. (1986*) Statistics, the exploration and analysis of data.* West Publishing Company, S:t Paul.

27 Ryan, T. P. (1990) Linear regression. In *Handbook of statistical methods for engineers and scientists*. Editor: Wadsworth, H. M. McGraw-Hill.

28 Zacharov, N.and Koivuniemi, K. (2001) Unravelling the perception of spatial sound reproduction: Analysis & preference mapping. Presented at *AES 111th Convention, New York*. Preprint 5423.

29 Cureton, E. E. and D'Agostino, R. B. (1983) *Factor Analysis – an applied approach*. Lawrence Erlbaum, New Jersey.

30 Bryman, A. and Cramer, D. (1994) *Quantitative data analysis for social scientists.* Routledge, London.

31 Bech, S. (1999) Methods for subjective evaluation of spatial characteristics of sound. In *Proceedings of the AES 16th International Conference on Spatial Sound Reproduction, 10–12 Apr*. pp 487-504. Audio Engineering Society.

32 IEC (1997) Draft IEC 60268-13. *Sound system equipment – part 13: listening test on loudspeakers*. International Electrotechnical Commission.

33 EBU (1990) Recommendation 562-3. *Subjective assessment of sound quality*. European Broadcasting Union.

34 Gabrielsson, A. and Sjögren A. (1979) Perceived sound quality of sound reproducing systems. *J. Acoust. Soc. Amer.* **65**, pp. 1019-1033

35 Toole, F. (1985) Subjective measurements of loudspeaker sound quality and listener performance. *J. Audio Engineering Society*. **33**, 1/2, pp 2-32.

36 Martin, G., Woszczyk, W., Corey, J. and Quesnel, R. (1999) Controlling phantom image focus in a multichannel reproduction system. Presented at *AES 107th Convention, New York*. Preprint 4996.

APPENDIX A

ATTRIBUTES TO ASSESS IN LISTENING TEST

GENERAL

Naturalness	How similar to a natural (i.e. not reproduced through e g loudspeakers) listening experience the sound as a whole sounds. Unnatural = low value. Natural = high value.
Presence	The experience of being in the same acoustical environment as the sound source, e g to be in the same room. Strong experience of presence = high value.
Preference	If the sound as a whole pleases you. If you think the sound as a whole sounds good. Try to disregard the *content* of the programme, i e do not assess genre of music or content of speech. Prefer the sound = high value.
Low frequency content	The level of low frequencies (the bass register). Low level ("less bass") = low value. High level ("much bass") = high value

SOUND SOURCE	*In some cases, more than one sound source (instrument/voice) occurs within the same sound excerpt. On the computer screen, you will be instructed which of these you should assess.*
Ensemble width	The perceived width/broadness of the ensemble, from its left flank to its right flank. The angle occupied by the ensemble. The meaning of "the ensemble" is all of the individual sound sources considered together. Does not necessarily indicate the known size of the source, e g one knows the size of a string quartet in reality, but the task to assess is how wide the sound from the string quartet is perceived. Disregard sounds coming from the sound source's environment, e g reverberation – only assess the width of the sound source. Narrow ensemble = low value. Wide ensemble = high value.
Individual source width	The perceived width of an individual sound source (an instrument or a voice). The angle occupied by this source. Does not necessarily indicate the known size of such a source, e g one knows the size of a piano in reality, but the task is to assess how wide the sound from the piano is perceived. Disregard sounds coming from the sound source's environment, e g reverberation – only assess the width of the sound source. Narrow sound source = low value. Wide sound source = high value.
Localisation	How easy it is to perceive a distinct location of the source – how easy it is to pinpoint the direction of the sound source. Its opposite (a low value) is when the source's position is hard to determine – a blurred position. Easy to determine the direction = high value.
Source distance	The perceived distance from the listener to the sound source. If several sources occur in the sound excerpt: assess the sound source perceived to be closest. Short distance/close = low value. Long distance = high value.
Source envelopment	The extent to which the sound source envelops/surrounds/exists around you. The feeling of being surrounded by the sound source. If several sound sources occur in the sound excerpt: assess the sound source perceived to be the most enveloping. Disregard sounds coming from the sound source's environment, e g reverberation – only assess the sound source. Low extent of envelopment = low value. High extent of envelopment = high value.

ROOM

Room width	The width/angle occupied by the sounds coming from the sound source's reflections in the room (the reverberation). Disregard the direct sound from the sound source. Narrow room = low value. Wide room = high value.
Room size	In cases where you perceive a room/hall, this denotes the relative size of that room. Large room = high value. If no room/hall is perceived, this should be assessed as zero.
Room sound level	The level of sounds generated in the room as a result of the sound source's action, e g reverberation – i e not extraneous disturbing sounds. Disregard the direct sound from the sound source. Weak room sounds = low value. Loud room sounds = high value.
Room envelopment	The extent to which the sound coming from the sound source's reflections in the room (the reverberation) envelops/surrounds/exists around you – i e not the sound source itself. The feeling of being surrounded by the reflected sound. Low extent of envelopment = low value. High extent of envelopment = high value.

APPENDIX B

ANOVA tables

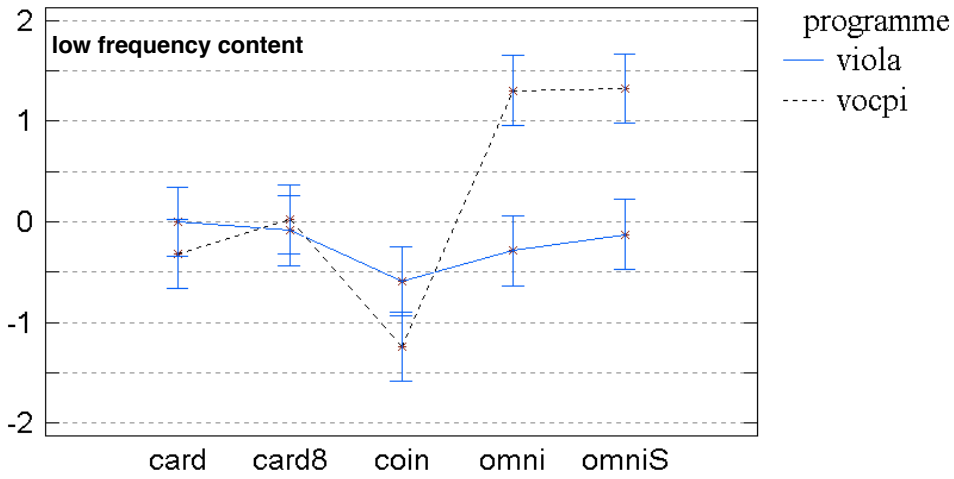
Attribute	Source	Sums of squares	Degrees of freedom	Mean square	*F*-ratio	*p*
lfc	rec_tech	47,426	4	11,8565	32,16	0,0000
	programme	7,60914	1	7,60914	20,64	0,0000
	rec_tech*programme	33,6637	4	8,41593	22,83	0,0000
	residual	55,3011	150	0,368674		
	total (corrected)	144	159			
nat	rec_tech	20,9894	4	5,24736	6,51	0,0001
	programme	0,664024	1	0,664024	0,82	0,3657
	rec_tech*programme	1,37681	4	0,344201	0,43	0,7891
	residual	120,97	150	0,806465		
	total (corrected)	144	159			
prf	rec_tech	31,4142	4	7,85355	12,68	0,0000
	programme	4,73991	1	4,73991	7,65	0,0064
	rec_tech*programme	14,9292	4	3,7323	6,03	0,0002
	residual	92,9167	150	0,619444		
	total (corrected)	144	159			
psc	rec_tech	9,23272	4	2,30818	2,98	0,0210
	programme	1,75154	1	1,75154	2,26	0,1346
	rec_tech*programme	16,9306	4	4,23265	5,47	0,0004
	residual	116,085	150	0,773901		
	total (corrected)	144	159			
dis	rec_tech	40,4886	4	10,1221	18,78	0,0000
	programme	6,16209	1	6,16209	11,44	0,0009
	rec_tech*programme	16,5227	4	4,13066	7,67	0,0000
	residual	80,8267	150	0,538845		
	total (corrected)	144	159			
ewd	rec_tech	30,0074	4	7,50185	14,46	0,0000
	programme	25,0675	1	25,0675	48,32	0,0000
	rec_tech*programme	11,1034	4	2,77586	5,35	0,0005
	residual	77,8217	150	0,518811		
	total (corrected)	144	159			
loc1	rec_tech	24,149	4	6,03724	10,71	0,0000
	programme	18,2988	1	18,2988	32,48	0,0000
	rec_tech*programme	17,0346	4	4,25866	7,56	0,0000
	residual	84,5176	150	0,563451		
	total (corrected)	144	159			
loc2	rec_tech	28,6102	4	7,15255	10,06	0,0000
	programme	0,580599	1	0,580599	0,82	0,3677
	rec_tech*programme	8,12697	4	2,03174	2,86	0,0256
	residual	106,682	150	0,711215		
	total (corrected)	144	159			

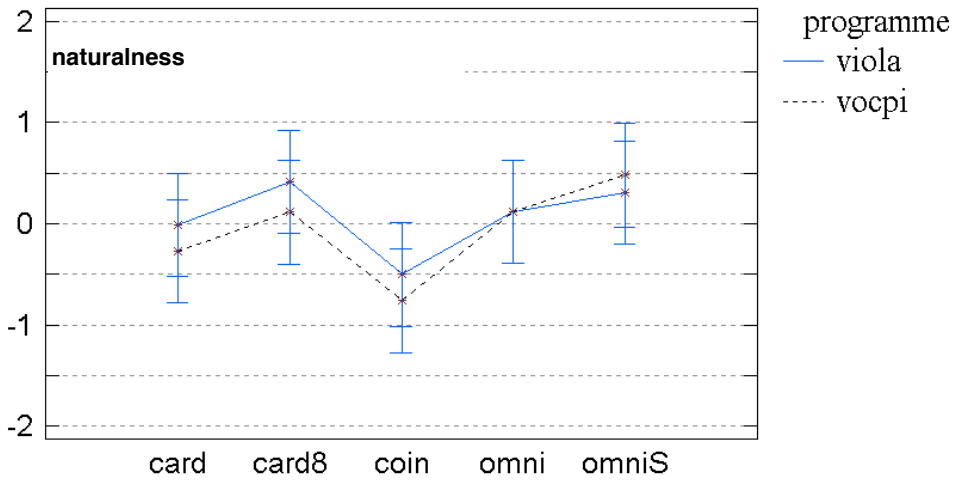
APPENDIX B – CONTINUED

Attribute	Source	Sums of squares	Degrees of freedom	Mean square	*F*-ratio	*p*
sev	rec_tech	42,341	4	10,5852	25,35	0,0000
	programme	3,55187	1	3,55187	8,51	0,0041
	rec_tech*programme	35,4669	4	8,86671	21,23	0,0000
	residual	62,6403	150	0,417602		
	total (corrected)	144	159			
swd1	rec_tech	53,3075	4	13,3269	36,34	0,0000
	programme	24,6714	1	24,6714	67,28	0,0000
	rec_tech*programme	11,014	4	2,75351	7,51	0,0000
	residual	55,0071	150	0,366714		
	total (corrected)	144	159			
swd2	rec_tech	29,3101	4	7,32751	10,39	0,0000
	programme	0,610204	1	0,610204	0,87	0,3538
	rec_tech*programme	8,29329	4	2,07332	2,94	0,0225
	residual	105,786	150	0,705243		
	total (corrected)	144	159			
rev	rec_tech	33,0167	4	8,25417	13,41	0,0000
	programme	2,06053	1	2,06053	3,35	0,0693
	rec_tech*programme	16,5854	4	4,14636	6,74	0,0001
	residual	92,3374	150	0,615582		
	total (corrected)	144	159			
rlv	rec_tech	61,0204	4	15,2551	29,6	0,0000
	programme	0,418762	1	0,418762	0,81	0,3688
	rec_tech*programme	5,2658	4	1,31645	2,55	0,0413
	residual	77,2951	150	0,515301		
	total (corrected)	144	159			
rsz	rec_tech	45,6542	4	11,4135	18,9	0,0000
	programme	0,687553	1	0,687553	1,14	0,2877
	rec_tech*programme	7,07942	4	1,76986	2,93	0,0228
	residual	90,5788	150	0,603859		
	total (corrected)	144	159			
rwd	rec_tech	47,0148	4	11,7537	19,81	0,0000
	programme	0,238089	1	0,238089	0,4	0,5274
	rec_tech*programme	7,7309	4	1,93273	3,26	0,0136
	residual	89,0162	150	0,593442		
	total (corrected)	144	159			

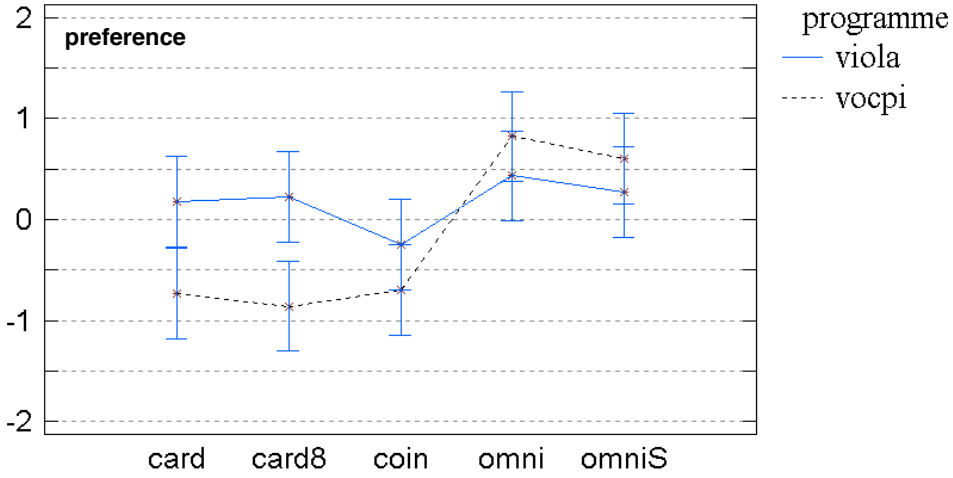
APPENDIX C

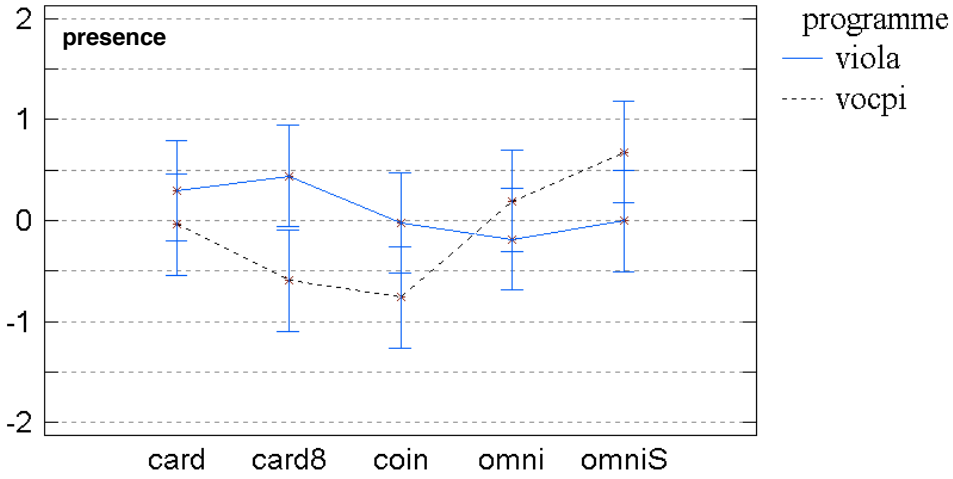
Interaction plots

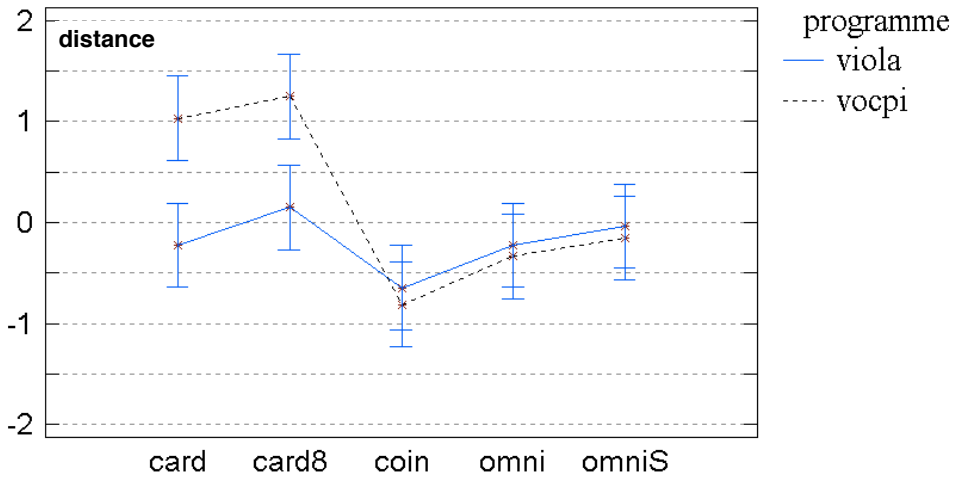


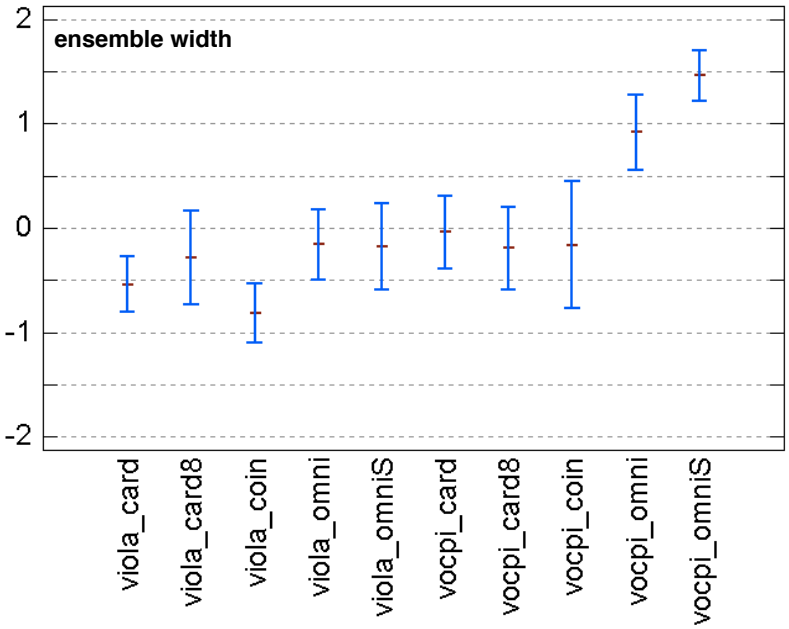


APPENDIX C – CONTINUED

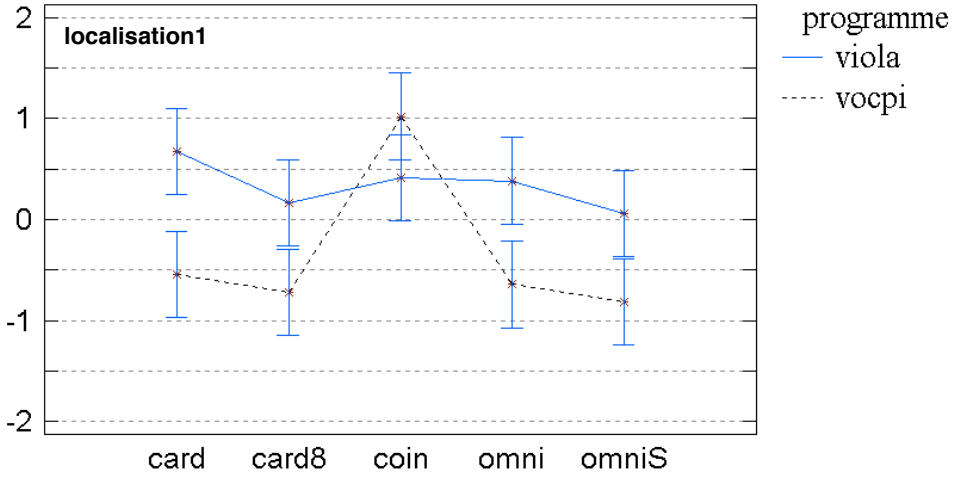


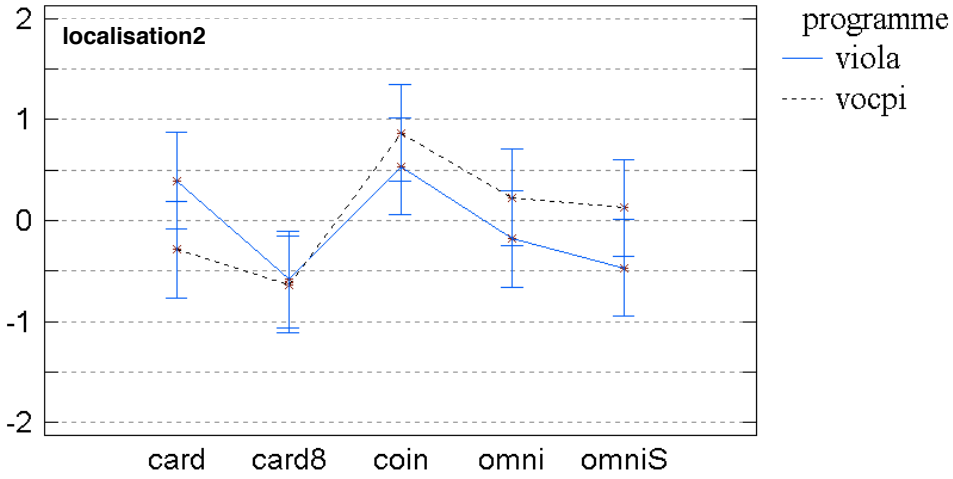


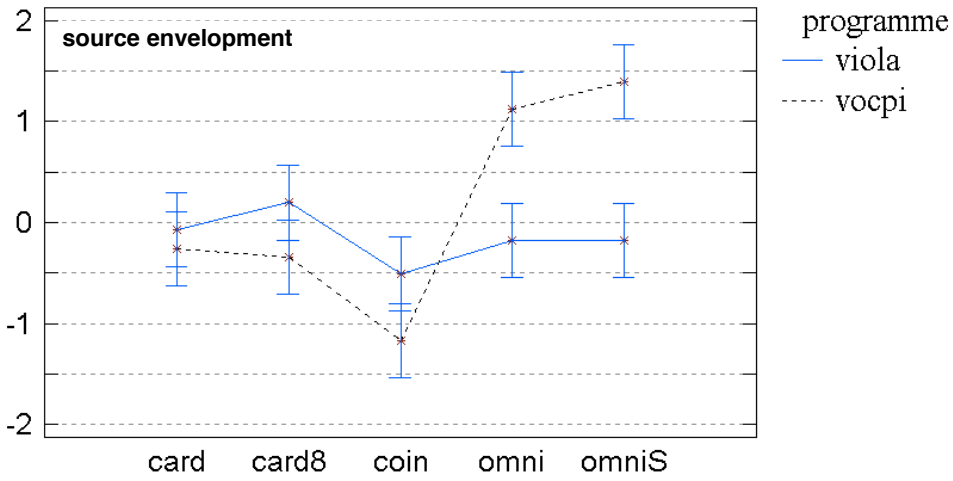


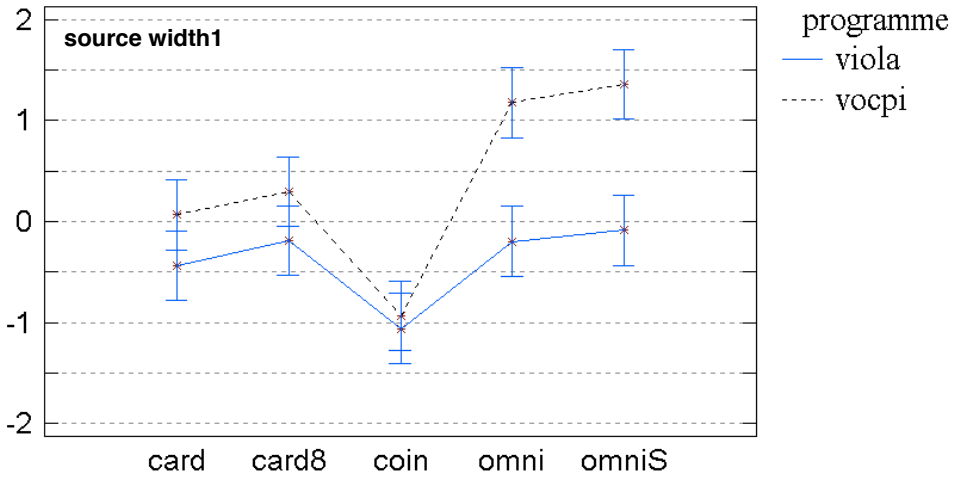


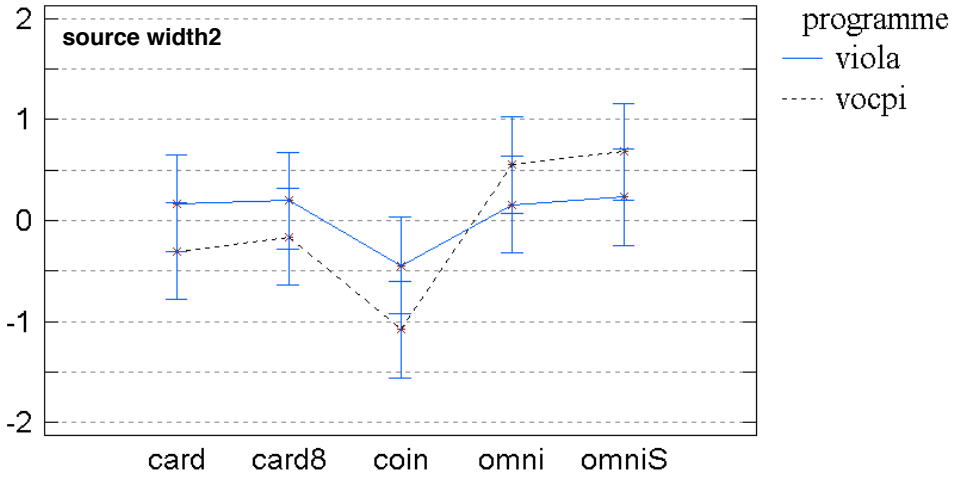
Intervals for ensemble width *are 95% confidence intervals calculated from the individual standard errors of each mean.*

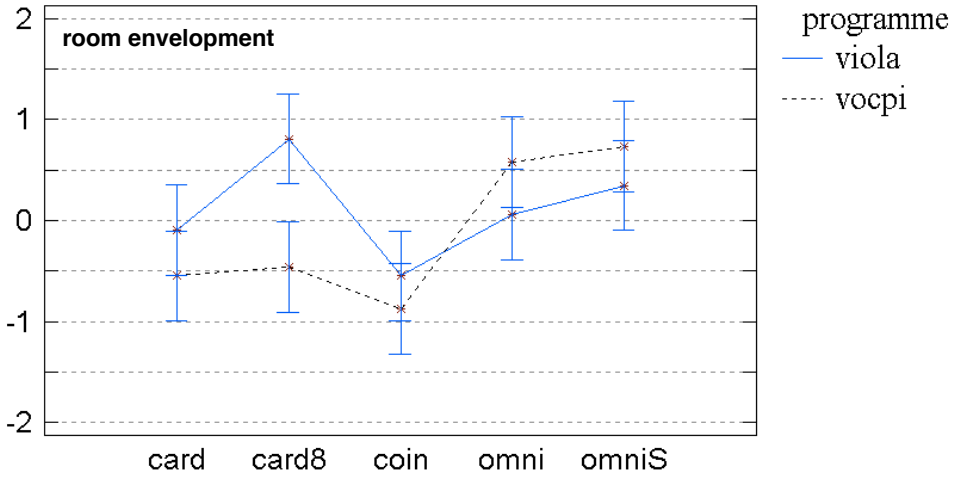

localisation1
2
1
0
-1
-2
card
card8
coin
omni
omniS
programme
viola
vocpi


localisation2
2
1
0
-1
-2
card
card8
coin
omni
omniS
programme
viola
vocpi

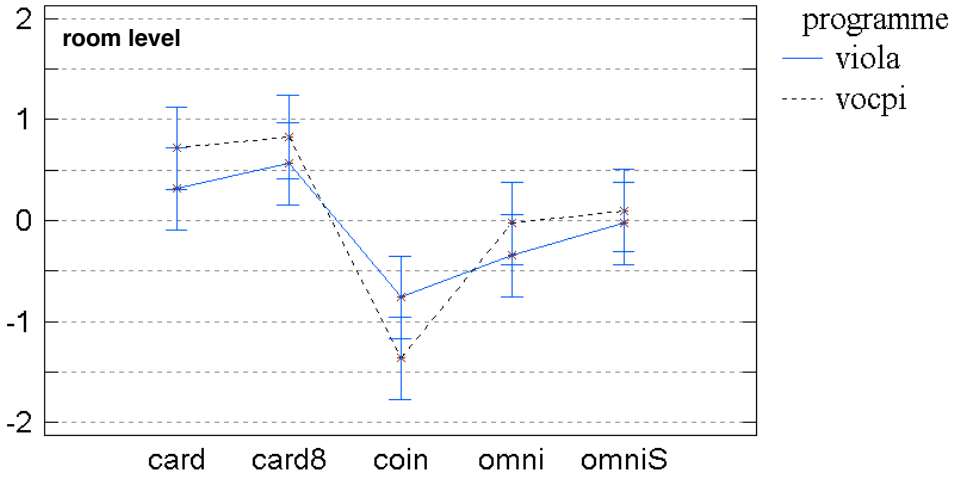

source envelopment
2
1
0
-1
-2
card
card8
coin
omni
omniS
programme
viola
vocpi

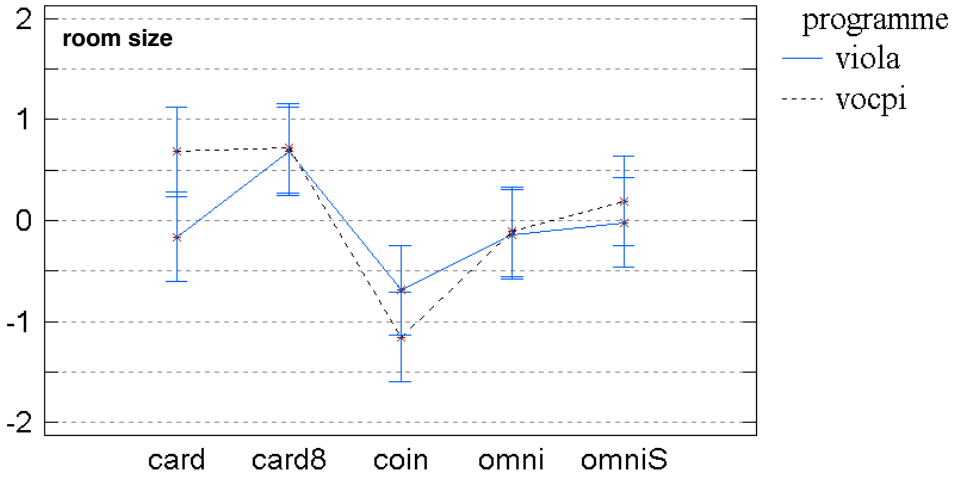

source width1
2
1
0
-1
-2
card
card8
coin
omni
omniS
programme
viola
vocpi


source width2
2
1
0
-1
-2
card
card8
coin
omni
omniS
programme
viola
vocpi

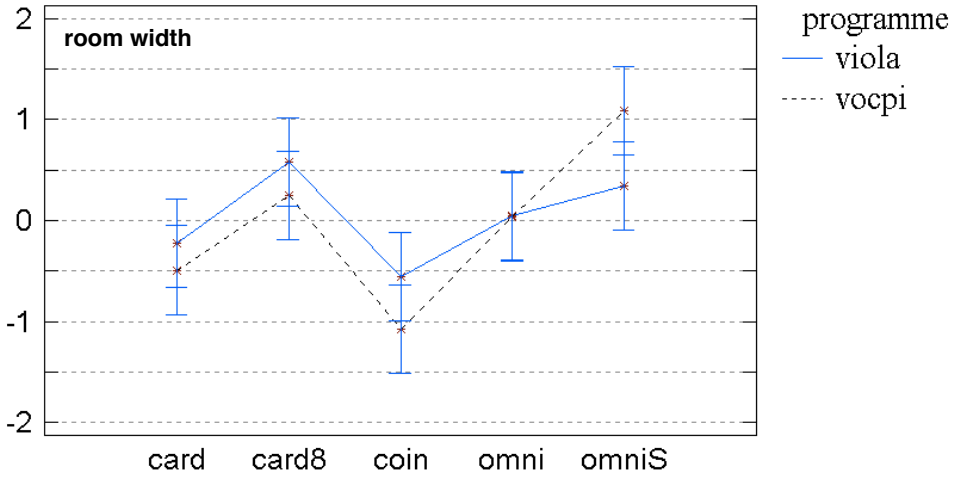

room envelopment
2
1
0
-1
-2
card
card8
coin
omni
omniS
programme
viola
vocpi

APPENDIX C – CONTINUED





APPENDIX C – CONTINUED



APPENDIX D

Correlations

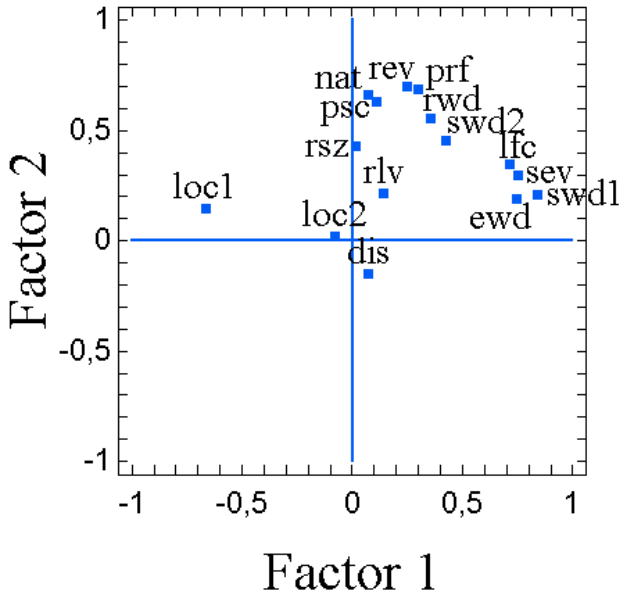
	lfc	nat	prf	psc	dis	ewd	loc1	loc2	sev	swd1	swd2	rev	rlv	rsz	rwd
lfc		0,276	0,367	0,282	0,072	0,454	-0,412	-0,081	0,650	0,643	0,401	0,451	0,275	0,256	0,425
nat	0,276		0,422	0,300	0,005	0,235	-0,051	0,004	0,253	0,225	0,243	0,313	0,123	0,287	0,304
prf	0,367	0,422		0,386	-0,221	0,313	-0,074	0,037	0,341	0,345	0,450	0,469	0,009	0,058	0,319
psc	0,282	0,300	0,386		0,003	0,212	-0,089	-0,037	0,279	0,221	0,229	0,376	0,137	0,175	0,344
dis	0,072	0,005	-0,221	0,003		-0,029	-0,451	-0,334	0,047	0,164	0,099	0,021	0,546	0,352	0,152
ewd	0,454	0,235	0,313	0,212	-0,029		-0,325	-0,059	0,497	0,657	0,334	0,283	0,094	0,131	0,332
loc1	-0,412	-0,051	-0,074	-0,089	-0,451	-0,325		0,326	-0,389	-0,548	-0,260	-0,140	-0,236	-0,200	-0,290
loc2	-0,081	0,004	0,037	-0,037	-0,334	-0,059	0,326		0,005	-0,143	-0,164	-0,142	-0,339	-0,193	-0,189
sev	0,650	0,253	0,341	0,279	0,047	0,497	-0,389	0,005		0,620	0,409	0,375	0,242	0,173	0,446
swd1	0,643	0,225	0,345	0,221	0,164	0,657	-0,548	-0,143	0,620		0,429	0,365	0,350	0,274	0,461
swd2	0,401	0,243	0,450	0,229	0,099	0,334	-0,260	-0,164	0,409	0,429		0,394	0,332	0,209	0,412
rev	0,451	0,313	0,469	0,376	0,021	0,283	-0,140	-0,142	0,375	0,365	0,394		0,193	0,358	0,492
rlv	0,275	0,123	0,009	0,137	0,546	0,094	-0,236	-0,339	0,242	0,350	0,332	0,193		0,467	0,399
rsz	0,256	0,287	0,058	0,175	0,352	0,131	-0,200	-0,193	0,173	0,274	0,209	0,358	0,467		0,395
rwd	0,425	0,304	0,319	0,344	0,152	0,332	-0,290	-0,189	0,446	0,461	0,412	0,492	0,399	0,395	

H1: Pearson product moment correlation coefficients

	lfc	nat	prf	psc	dis	ewd	loc1	loc2	sev	swd1	swd2	rev	rlv	rsz	rwd
lfc		0,0004	0	0,0003	0,3693	0	0	0,3117	0	0	0	0	0,0004	0,0011	0
nat	0,0004		0	0,0001	0,9485	0,0028	0,5215	0,9637	0,0012	0,0042	0,002	0,0001	0,1226	0,0002	0,0001
prf	0	0		0	0,0049	0,0001	0,3524	0,6466	0	0	0	0	0,9099	0,4657	0
psc	0,0003	0,0001	0		0,9669	0,007	0,2611	0,6447	0,0004	0,005	0,0035	0	0,0834	0,0265	0
dis	0,3693	0,9485	0,0049	0,9669		0,7183	0	0	0,5582	0,0379	0,2154	0,7892	0	0	0,0551
ewd	0	0,0028	0,0001	0,007	0,7183		0	0,4603	0	0	0	0,0003	0,2366	0,0999	0
loc1	0	0,5215	0,3524	0,2611	0	0		0	0	0	0,0009	0,078	0,0026	0,0111	0,0002
loc2	0,3117	0,9637	0,6466	0,6447	0	0,4603	0		0,9483	0,0707	0,0377	0,0731	0	0,0147	0,0168
sev	0	0,0012	0	0,0004	0,5582	0	0	0,9483		0	0	0	0,0021	0,0286	0
swd1	0	0,0042	0	0,005	0,0379	0	0	0,0707	0		0	0	0	0,0004	0
swd2	0	0,002	0	0,0035	0,2154	0	0,0009	0,0377	0	0		0	0	0,0081	0
rev	0	0,0001	0	0	0,7892	0,0003	0,078	0,0731	0	0	0		0,0143	0	0
rlv	0,0004	0,1226	0,9099	0,0834	0	0,2366	0,0026	0	0,0021	0	0	0,0143		0	0
rsz	0,0011	0,0002	0,4657	0,0265	0	0,0999	0,0111	0,0147	0,0286	0,0004	0,0081	0	0		0
rwd	0	0,0001	0	0	0,0551	0	0,0002	0,0168	0	0	0	0	0	0	

H2: p-values for non-correlation. A single "0" denotes p<0.00005

APPENDIX E



Factor loadings – all attributes