Audio Engineering Society

Convention Paper 7783

Presented at the 126th Convention
2009 May 7–10 Munich, Germany

Comparing three methods for sound quality evaluation with respect to speed and accuracy

Florian Wickelmaier[1], Nora Umbach[1], Konstantin Sering[1], and Sylvain Choisel[2]

[1] *Dept. of Psychology, University of Tübingen, Tübingen, Germany*

[2] *Bang & Olufsen A/S, Struer, Denmark; now at Philips Consumer Lifestyle, Leuven, Belgium*

Correspondence should be addressed to Florian Wickelmaier (florian.wickelmaier@uni-tuebingen.de)

ABSTRACT

The goal of the present study was to compare three response-collection methods that may be used in sound quality evaluation. To this end, 52 listeners took part in an experiment where they assessed the audio quality of musical excerpts and six processed versions thereof. For different types of program material, participants performed (a) a direct ranking of the seven sound samples, (b) pairwise comparisons and (c) a novel procedure, called ranking by elimination. The latter requires subjects on each trial to eliminate the least preferred sound; the elimination continues until only the sample with the highest audio quality is left. The methods are compared with respect to the resulting ranking/scaling and the time required to obtain results.

1. INTRODUCTION

Several psychometric methods exist to measure listeners' preferences among different systems, such as loudspeakers, upmix algorithms and audio codecs. The de-facto standard in sound quality evaluation are direct scaling procedures, as described in recommendations of the International Telecommunications Union (ITU-R BS.1116-1 [5], ITU-R BS.1534-1 [6], ITU-T P800 [7]). In direct scaling, participants are required to assign a number or a verbal label to the sounds that are supposed to reflect the magnitude of the sensation evoked by these sounds. Direct scaling is widely used, but it relies on many implicit and untested assumptions and is known to give rise to multiple biases [14].

An alternative approach to the quantification of auditory attributes is indirect scaling. The motivation of indirect scaling is that global and specific auditory attributes, like audio quality, preference, spaciousness, brightness, are latent psychological vari-

ables. Their magnitudes cannot be observed directly but need to be inferred from observable behavior. An application of indirect scaling in the context of multichannel sound is described in [2], where paired comparisons and subsequent modeling of the data have been used in order to quantify overall preference as well as more specific attributes. Paired-comparison scaling has advantages of reducing the subject's tasks to simple pairwise judgments, avoids biases linked to scale usage, and enables the experimenter to perform consistency checks related to unidimensionality of the attribute being measured. The main disadvantage of the paired-comparison method is the time required, making it expensive in many practical situations.

A possible trade-off between the speed of a listening test and its accuracy is the use of ranking procedures, where listeners have access to all stimuli (or possibly only a subset) and rank them, for example, according to their preference. This may eradicate scaling biases, and a numerical scale can either be directly derived from the ranks across repetitions or subjects, or, preferably, by subsequent modeling of the ranking data. In practice, however, when performing such a ranking, some listeners spend a long time doing pairwise comparisons, and often reconsider and alter their rankings. Therefore, a new ranking procedure is proposed, where the task is to identify the least preferred stimuli, which will subsequently (and irrevocably) be eliminated from the list. The task is repeated until only one stimulus is left. This is thought to increase the speed of the listening test and ease the subject's task, thereby reducing frustration.

It was the goal of the present study, to compare three response collection methods, paired comparisons, a traditional ranking procedure and ranking by elimination with respect to the scaling of the sounds, and to the time required to complete the respective task. We hypothesized that ranking by elimination would be faster than conventional ranking while being as accurate as the pairwise comparisons. Furthermore, the present experiment makes it possible to investigate the individual strategies listeners might have when assessing the sounds, such as changing ranks or repeated listening to certain sound samples. It was another goal of the study to reveal such idiosyncratic strategies if they occur.

2. METHODS

2.1. Subjects

The sample consisted of 52 participants (41 female), most of them undergraduate psychology students who received course credits. They were between 18 and 44 years old ($M = 23.7$). None of the subjects reported any hearing problems.

2.2. Apparatus

The experiment was conducted in a sound-attenuating booth. The sounds were played back by a personal computer equipped with an external sound card (RME Hammerfall DSP Multiface II) connected to a headphone amplifier (Behringer Powerplay PRO-8 HA8000), and were delivered by a pair of headphones (Beyerdynamic DT990 Pro). Stimulus presentation and response collection was controlled by custom made software written in Python. Listeners entered their responses by clicking on buttons displayed on a 19-in LCD monitor (EIZO FlexScan S1921-SE) using an optical mouse.

2.3. Program material and stimuli

Three musical excerpts were selected from commercially available audio CDs (Table 1), recorded into WAV format at 44.1 kHz sampling frequency and 16-bit resolution, and carefully cut to include a musical phrase of about 5 s duration. None of the excerpts included vocals. The material was selected to be representative of changes in audio quality that are induced by compression algorithms, the castanets excerpt being especially prone to coding artifacts, the remaining two excerpts, constituting examples of bandwidths and content typical of pop and classical music. These unprocessed files were encoded using MP3 and OGG VORBIS codecs as implemented in software (Audacity version 1.3.2 using the LAME encoder version 3.97 for MP3 encoding). For each type of program material, three MP3 encoded versions were generated employing constant bit rates of 128, 96, and 64 kbps. In addition, three OGG encoded versions were created at quality settings q4, q2, and q0, corresponding to nominal bit rates of 128, 96, and 64 kbps, respectively.

The encoded versions were transformed into WAV files, and 20 ms linear rise and fall times were applied to prevent clicks. Within each type of program material, the RMS amplitudes of the six encoded versions were aligned with the unprocessed version, in

Table 1: List of musical program material.

Disc	Title	Track	Time
AES: Perceptual Audio Coders	Castanets Original	01	0'00–0'04
Funkaholish: By Your Side	Vincenzo's Japanese Limousine	06	0'14–0'19
Bach: Christmas Oratorio BWV 248 – conducted by Ludwig Güttler	Coro "Jauchzet, frohlocket"	01	0'17–0'22

order to prevent loudness differences as much as possible. The play-back levels were adjusted beforehand to a comfortable level by the experimenters. The unweighted sound-pressure levels as measured with an artificial ear and a sound-level meter (CEL-275) were 60.4 dB SPL for the castanets, 76.2 dB SPL for the pop music, and 68.3 dB SPL for the classical music.

2.4. Procedure

The participants completed three blocks in which they judged the three program materials using one of three response collection methods. In the first block, subjects did all 21 pairwise comparisons for each program material. Their task was on each trial to choose the sound that had the higher audio quality. Within a pair, all sounds were played back only once. A completely balanced paired-comparison design [4] ensured that each compression setting appeared equally often in position one and two. The within-pair order of the sounds was balanced across subjects. The order of the 21 pairs was randomized. The three musical excerpts were completed successively and presented in random order. After the 21 comparisons, there was a short self-paced break. The first block started with three warm-up trials that were discarded from analysis.

In the second block, half of the listeners performed a regular ranking task for the three musical excerpts, the other half of the subjects completed a ranking by repeatedly eliminating the sound having the lowest quality, as will be described below. On each trial in the regular ranking, participants were presented with all seven compression settings represented as buttons on the screen and asked to rank them with respect to audio quality. The ranks were entered by clicking on buttons labeled from 1 to 7 below the corresponding sound button (see Figures 4 and 5 for a schematical representation of the response interface). Subjects were free to replay each sound as often as they wanted. They were also free to alter their rankings as often as they desired, until they confirmed their final ranking. No time restrictions were given. After a self-paced break, the ranking continued for the remaining two musical excerpts. The order of the compression settings on the screen was random, as was the order of the excerpts within the block. In order to familiarize the participants with the response interface, the ranking block started with a short warm-up trial.

The response interface for the ranking by elimination was similar to the regular ranking procedure, but this time only the highest rank (lowest quality) could be clicked which subsequently disappeared from the screen. The elimination continued until only one compression setting, the one with the highest audio quality, was left. Listeners were able to replay the available sounds as often as needed, but it was not possible to alter a decision once made. As in the regular ranking, no time restrictions were imposed, and subjects completed each program material in turn after a self-paced break. The order of the compression settings and of the excerpts were randomized. At the beginning of the block there was a short warm-up trial.

Half of the participants first performed the ranking task and then ranking by elimination, the other half did the reversed order. Between the three blocks (response collection methods) there was a longer break of about 5 min. The total duration of the experiment was about 50 min.

2.5. Analysis of the choices and rankings

Pairwise judgments were aggregated across subjects in order to estimate pairwise choice probabilities, P_{xy}, of choosing sound x over sound y with respect to its audio quality. In order to test whether the judgments fulfill the structural requirements necessary for a scale to be derived, the number of violations of the stochastic transitivities were counted. The weak (WST), moderate (MST), and strong

(SST) stochastic transitivities imply that if $P_{xy} \geq 0.5$ and $P_{yz} \geq 0.5$, then

$$P_{xz} \geq \begin{cases} 0.5 & \text{(WST)} \\ \min\{P_{xy}, P_{yz}\} & \text{(MST)} \\ \max\{P_{xy}, P_{yz}\} & \text{(SST)} \end{cases} \quad (1)$$

for all sounds x, y and z. Whenever the premise holds, but the implication in Equation 1 does not hold (for any permutation of the triple x, y, z), a transitivity violation is observed. While WST is a necessary condition for an ordinal scale to be derived, choice models that imply higher scale levels require MST to hold or even SST.

Two kinds of probabilistic choice models were considered for representing the choice frequencies. The rationale of these models is that they connect observable quantities (the choice probabilities) with a latent psychological attribute (in this application: audio quality), thus allowing one to estimate the magnitudes of this latent attribute. First, the Bradley-Terry-Luce (BTL) model [1, 8], which predicts P_{xy} as a function of parameters associated with each sound

$$P_{xy} = \frac{u(x)}{u(x) + u(y)}, \quad (2)$$

where $u(\cdot)$ is a ratio scale of the criterion. The BTL model requires SST.

When the BTL model did not account for the data, the more general elimination-by-aspects (EBA) model [10, 11] was attempted to fit. According to EBA, one sound is chosen over a second one because of a certain *aspect* which belongs to the first but not to the second sound. EBA predicts P_{xy} by

$$P_{xy} = \frac{\sum_{\alpha \in x' \setminus y'} u(\alpha)}{\sum_{\alpha \in x' \setminus y'} u(\alpha) + \sum_{\beta \in y' \setminus x'} u(\beta)}, \quad (3)$$

where $\alpha, \beta, \ldots$ are the aspects (or features) of the sounds, and $x' \setminus y'$ denotes the set of aspects belonging to sound x but not to sound y. As for the BTL model, $u(\cdot)$ is a ratio scale of the criterion, but the EBA model only required MST.

The ranking data were aggregated across participants in oder to estimate the probability of each ranking. In those cases where the BTL model provided an adequate fit, the Mallows-Bradley-Terry (MBT) model [9, 3] was applied. This model predicts the probability of each ranking R_m by

$$P(R_m) = \frac{\prod_{i=1}^{t} u_i^{t-r_{im}}}{\sum_{m=1}^{t!} \prod_{i=1}^{t} u_i^{t-r_{im}}}, \quad (4)$$

where t is the number of sounds. This is, in essence, the product of all pairwise probabilities consistent with a given ranking. For example, when $t = 7$, $R_1 = (1, 2, 3, 4, 5, 6, 7), \ldots, R_{5040} = (7, 6, 5, 4, 3, 2, 1)$. Since only few of these possible rankings can actually be observed with $n = 52$ subjects, a goodness-of-fit test is not possible. It is assumed that whenever the BTL model fits the corresponding paired-comparison data, the MBT model would fit the rankings, and vice versa. As for the choice models, the u parameters of the MBT model are a ratio scale.

Fitting and testing of the models was done using software programmed in the R environment (`www.r-project.org`). Portions of this software are described in [13, 2, 12].

3. RESULTS

3.1. Scaling audio quality

The choice proportions for the three musical excerpts are displayed in Table 2. The most extreme proportions occur with the castanets, for the remaining excerpts proportions are closer to 0.5.

Table 3 shows the number of violations of the weak, moderate, and strong stochastic transitivities. There were only few violations of WST, which suggests that an ordinal scale can be derived from the judgments for each of the excerpts. These ordinal scales are given by the order of the columns in Table 2. For castanets, MST and SST violations are relatively few, for the remaining two excerpts they seem to be more substantial.

The goodness-of-fit tests in Table 3 indicate that the BTL model holds for castanets and pop music, but does not hold for the classical excerpt. This suggests

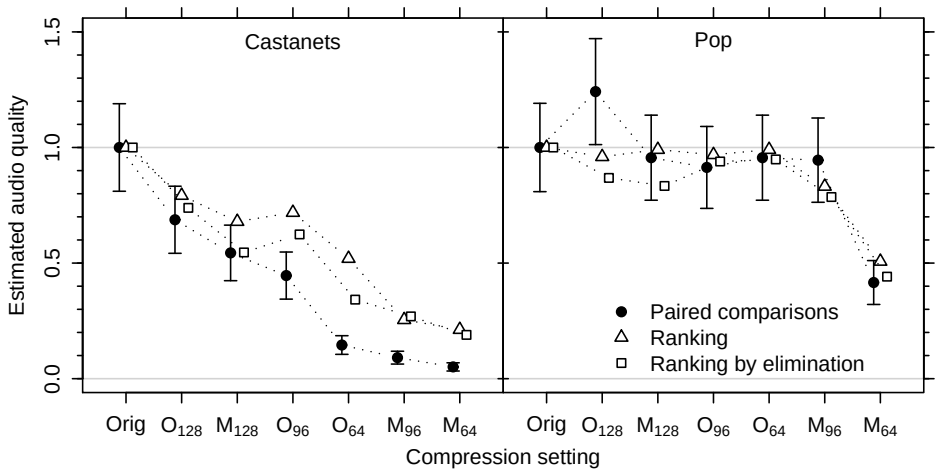

Fig. 1: Audio quality scale values derived from three response-collection methods: pairwise comparisons (parameter estimates of the Bradley-Terry-Luce model with approximate 95% confidence intervals), ranking and ranking by elimination (parameter estimates of the Mallows-Bradley-Terry model).

for the former two excerpts that the observed transitivity violations are not systematic, for the latter that at least the SST violations are more frequent than expected by chance. In order to further explore the structure of the judgments for the classical excerpt, two simple EBA models were attempted to fit: the first had two additional aspects, one for MP3 and one for OGG codec type; the second had one aspect for each of the three bit rate settings; neither could account for the data ($\chi^2(13) = 28.36$, $p = .008$ and $\chi^2(12) = 23.95$, $p = .021$).

To verify that the listeners could perceive any difference at all between the sounds, further tests were conducted. For the castanets and the pop excerpts, the respective BTL models were compared to a null model that predicts $P_{ij} = 0.5$ for all pairs. The null model could be rejected (castanets: $\chi^2(6) = 434.73$, $p < .001$; pop: $\chi^2(6) = 59.54$, $p < .001$). For the classical music, since the BTL model failed to hold, the null model was compared to the saturated model that fits the data perfectly. Here too, the null model was rejected ($\chi^2(21) = 56.63$, $p < .001$), indicating that the subjects perceived differences between some compression settings, even for the classical music.

Figure 1 shows the scale values derived for the castanets and the pop excerpts. The scale values are the parameter estimates of the Bradley-Terry-Luce model applied to the paired-comparison data and the parameter estimates of the Mallows-Bradley-Terry model applied to the ranking data. Scale values are normalized such that $u(\text{Orig}) := 1$. For castanets, the perceived audio quality spans a considerable range: Taking advantage of the ratio-scale property of the u-scale, since $\hat{u}(\text{M}_{64}) = 0.05$, the uncompressed version has an audio quality about 20 times higher than the MP3 at 64 kbps for the paired comparisons.

Overall the scale values derived from the three response collection methods seem to be in good agreement. The mean absolute deviation, $\frac{1}{t}\sum_i^t |\hat{u}_i - \hat{v}_i|$, between BTL and MBT scale values derived from regular ranking is 0.17 for castanets and 0.09 for pop. For ranking by elimination it is 0.11 for castanets and 0.10 for pop. There is a tendency, however, for the ranking methods to produce less pronounced differences than the paired comparisons. For the castanets, the uncompressed version has an audio quality only about five times higher than the MP3 at 64 kbps for both ranking procedures (as compared to 20 times higher for paired comparisons). Testing the MBT model against the null model (indifference test) yields significant results for ranking (castanets: $\chi^2(6) = 204.29, p < .001$; pop: $\chi^2(6) = 60.61, p < .001$) and for ranking by elimination (castanets: $\chi^2(6) = 213.01, p < .001$; pop: $\chi^2(6) = 71.92, p < .001$).

Table 2: Choice proportions, $n = 52$ for each pair.

Castanets							
	Orig	O_{128}	O_{96}	M_{128}	O_{64}	M_{96}	M_{64}
Orig	–	.54	.73	.73	.81	.90	.96
O_{128}	.46	–	.56	.50	.83	.92	.94
O_{96}	.27	.44	–	.52	.77	.75	.88
M_{128}	.27	.50	.48	–	.85	.90	.90
O_{64}	.19	.17	.23	.15	–	.65	.71
M_{96}	.10	.08	.25	.10	.35	–	.67
M_{64}	.04	.06	.12	.10	.29	.33	–
Pop							
	O_{128}	M_{96}	M_{128}	Orig	O_{64}	O_{96}	M_{64}
O_{128}	–	.67	.60	.50	.58	.44	.79
M_{96}	.33	–	.42	.58	.54	.52	.73
M_{128}	.40	.58	–	.54	.38	.50	.73
Orig	.50	.42	.46	–	.62	.60	.62
O_{64}	.42	.46	.62	.38	–	.52	.73
O_{96}	.56	.48	.50	.40	.48	–	.63
M_{64}	.21	.27	.27	.38	.27	.37	–
Classic							
	Orig	M_{96}	O_{96}	O_{128}	O_{64}	M_{128}	M_{64}
Orig	–	.58	.56	.54	.52	.52	.60
M_{96}	.42	–	.67	.60	.60	.60	.56
O_{96}	.44	.33	–	.54	.56	.29	.56
O_{128}	.46	.40	.46	–	.67	.62	.73
O_{64}	.48	.40	.44	.33	–	.58	.58
M_{128}	.48	.40	.71	.38	.42	–	.65
M_{64}	.40	.44	.44	.27	.42	.35	–

For the classical excerpt, since neither the BTL model nor a-priori plausible EBA models fitted the data, only ordinal scales were derived. Based on the median rankings, the rank order (from best to worst) of the compression settings for regular ranking is (Orig, O_{128}, M_{128}, M_{96}, O_{96}, O_{64}, M_{64}); for ranking by elimination it is (Orig, O_{128}, O_{96}, M_{128}, M_{96}, O_{64}, M_{64}). At first glance these rank orders seem to be rather different from the rank order obtained by paired comparisons (given in Table 2). Note, however, that the rank orders do not contain information as to which compression settings are significantly different; only for the paired comparisons was the indifference rejected by a formal statistical test as described before.

3.2. Time required

Figure 2 shows the time participants needed to complete an evaluation of the compression settings using a given method. It is obvious, that paired comparisons take longest on average for all types of program material. It can also be seen that the variance is smallest for the paired comparisons, and that it is very large for the ranking methods. For the castanets, some listeners only take about 50 s to complete the ranking, while some of them take longer than 300 s. For the pop and classical excerpts, some participants take even longer than 600 s for the ranking or ranking by elimination.

Table 3: Number violations of the weak (WST), moderate (MST) and strong (SST) stochastic transitivity out of 35 possible tests, and goodness-of-fit tests of the BTL model for each excerpt.

	WST	MST	SST	$\chi^2(15)$	p
Castanets	0	1	8	13.18	.588
Pop	4	13	20	21.01	.137
Classic	2	12	29	27.03	.029

There is a tendency that the ranking by elimination procedure is slightly faster than the regular ranking; this is significant for castanets ($t(51) = 2.26$, $p = .028$), but not significant for the pop ($t(51) = 1.31$, $p = .196$) and the classical excerpt ($t(51) = 0.62$, $p = .538$).

Figure 3 shows the time difference between clicks on ranking buttons for the ranking by elimination method. For the pop and classical excerpts this time difference is decreasing, indicating that subjects become faster the more sounds have been eliminated. Thus, although the judgments presumably become more difficult in the end, the fact that only fewer stimuli need to be compared leads to quicker decisions. The general pattern of faster responses at the end of the trial can also be seen for the castanets. Apparently, the rather obvious quality degradations induced by the low-bit-rate MP3 versions make the first judgments faster than for the remaining excerpts, leading to an inverse u-shaped relationship between the progress of procedure and time difference.

3.3. Response behavior

Table 4 displays the average number of times participants clicked in order to replay a certain compression setting. For the castanets, the number of sound

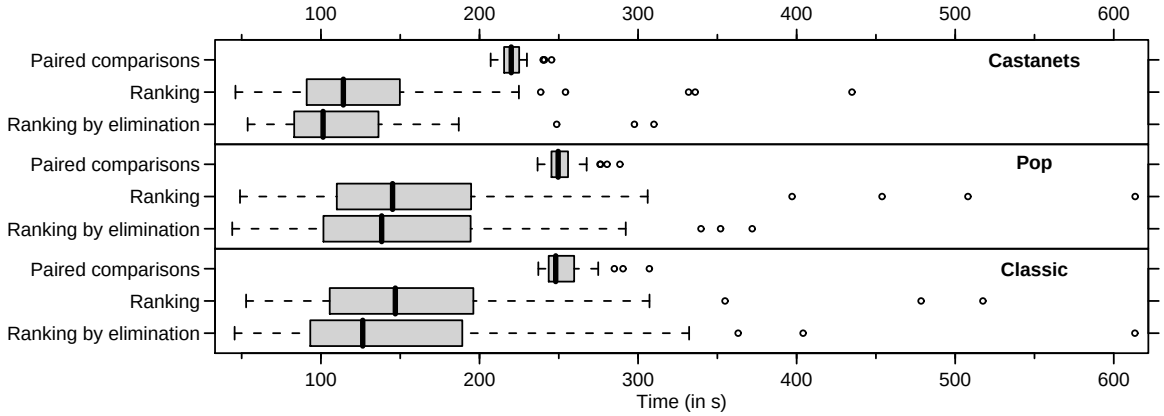

Fig. 2: Boxplots of the time required for an evaluation of the compression settings using one of the three methods.

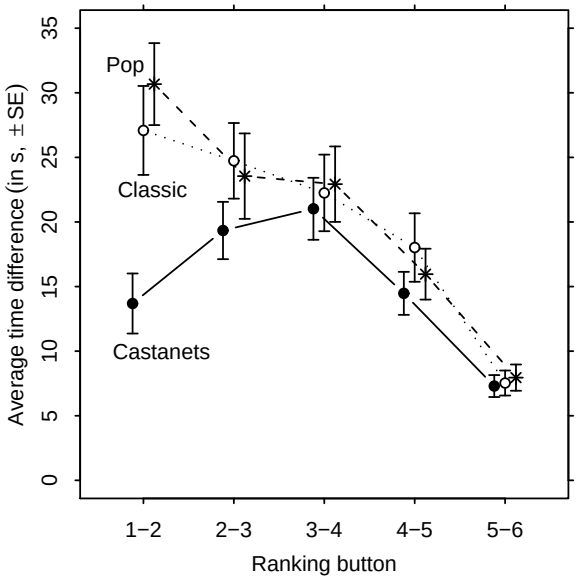

Fig. 3: Time difference between clicks on ranking buttons for ranking by elimination.

button clicks is lower for settings that were judged to be of lower quality (MP3 at 64 and 96 kbps, respectively), while higher-quality formats were played back more often. For the pop excerpt, the MP3 at 64 kbps was clicked on average 1.5 times (ranking) and 0.9 times (ranking by elimination) while the remaining settings were played back more often. For the classical music, all compression settings were clicked about equally often in the ranking procedures, only the MP3 at 64 kbps being clicked less frequently in ranking by elimination.

For the castanets, the number of sound button clicks was lower than for the remaining two excerpts. Over all compression settings, sound button clicks were slightly less in the ranking procedure than in ranking by elimination; this was, however, not significant for the castanets ($t(51) = 0.92$, $p = .361$). For the pop and classical excerpts, the number of sound button clicks in the two methods was about the same (pop: $t(51) = -0.17$, $p = .865$; classic: $t(51) = -0.58$, $p = .564$).

There were large individual differences in how often subjects played back the sounds and in how often they adjusted their rankings. Two examples are given in Figures 4 and 5. The figures schematically illustrate the response interface that was displayed on the screen. There was a top row of sound buttons (denoted by +) that allowed for repeated play-back; below the sound buttons, there were additional seven rows of buttons labeled from 1 to 7 for the ranking. The small black numbers in Figures 4 and 5 are event counts. For example, in Figure 4 the listener first played back the MP3 at 96 kbps, then she assigned it rank 1; subsequently, she listened to the OGG at 64 kbps and ranked it to be first; then she adjusted the rank for the MP3 at 96 kbps to be second, and so on. There were 20 clicks on the ranking buttons, indicating that this participant reconsidered and altered previous rankings several times throughout the

Table 4: Average number of sound button clicks for ranking (Rnk) and ranking by elimination (Elm); the difference between the two numbers is denoted by Δ.

	Castanets			Pop			Classic		
	Rnk	Elm	Δ	Rnk	Elm	Δ	Rnk	Elm	Δ
Orig	2.6	2.6	−0.02	2.7	3.2	−0.58	2.8	3.0	−0.23
M_{128}	2.5	2.4	0.12	2.5	2.5	−0.04	2.6	3.0	−0.37
M_{96}	1.4	1.2	0.23	2.7	2.4	0.23	2.4	2.6	−0.12
M_{64}	1.2	0.9	0.35	1.5	0.9	0.65	2.4	1.8	0.62
O_{128}	2.9	2.9	0.02	3.0	2.8	0.12	2.6	3.1	−0.54
O_{96}	2.8	2.5	0.25	2.8	3.1	−0.33	2.7	3.1	−0.46
O_{64}	1.8	1.5	0.21	2.8	3.1	−0.35	2.3	2.3	−0.02
Sum	15.3	14.1	1.15	17.8	18.1	−0.29	17.8	18.9	−1.12

procedure.

Figure 5 displays the response behavior of a different subject. In contrast to the previous one, this listener does not alter her rankings once they have been assigned, but she frequently clicks on the play-back buttons (37 times). Note that together with the forced play-back at the beginning of the ranking trial, the sounds were played back 44 times. In a complete paired comparison, the sounds are replayed 42 times in total.

4. DISCUSSION

The present paper compares three methods for sound quality evaluation with respect to speed and

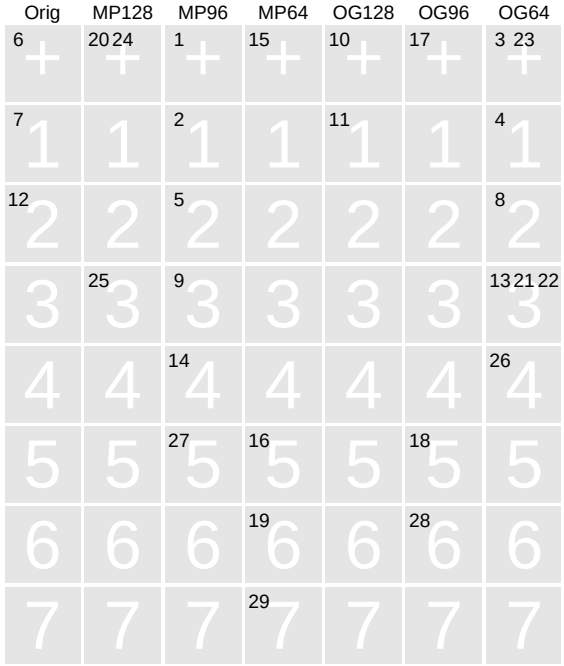

Fig. 4: A listener's (id37) response behavior for the pop excerpts as measured by button clicks. The black numbers are event counts.

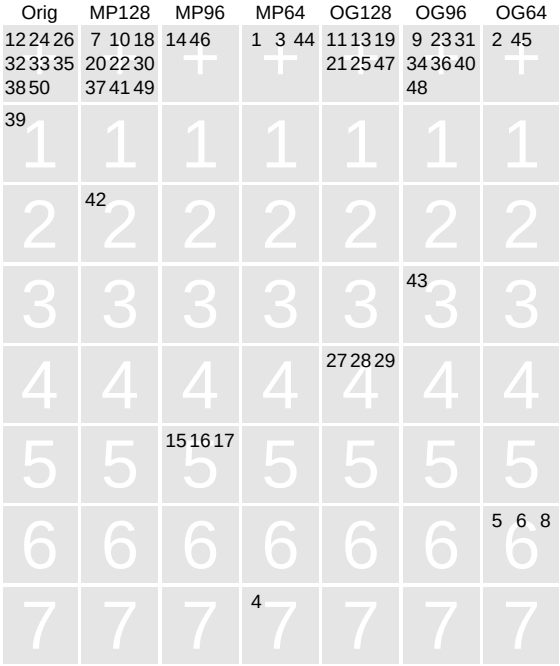

Fig. 5: A listener's (id13) response behavior for the castanets excerpts as measured by button clicks. The black numbers are event counts.

accuracy. Subjects completed all three methods in a within-subject design. They had to evaluate three musical excerpts that have been encoded by MP3 and OGG VORBIS codecs at (nominal) bit rates of 128, 96, and 64 kbps with respect to audio quality. The seventh version was the original uncompressed excerpt.

For the castanets results show that the data for pairwise comparisons were consistent with the restrictions of the BTL model. Therefore, a ratio scale of perceived quality was derived from the data with no significant violations of stochastic transitivity. Participants perceived the original version of highest quality and the MP3 with a bit rate of 64 kbps of lowest quality. The BTL model also held for paired-comparison data of the pop excerpt. This time the compression settings were of comparable audio quality, except for the MP3 at 64 kbps for which quality dropped to about 40% (Figure 1). For paired-comparison data of the classical excerpt, violations of moderate and strong stochastic transitivity seemed to be systematic and no BTL or ad-hoc EBA model could be fitted. For the derived ordinal scale, the original excerpt was again perceived of highest quality and the 64-kbps MP3 as lowest quality. For all three excerpts indifference tests indicated that subjects perceived differences between some compression settings, even for the classical excerpt.

For the other two methods, ranking and ranking by elimination, the MBT model was fitted to the data of the castanets and the pop excerpt. Since the BTL model did not fit the paired-comparison data of the classical excerpt, it was not attempted to fit an MBT model to the ranking data. The MBT scale values for the other two excerpts show the same pattern as the BTL scale values for the paired-comparison data (Figure 1). Overall, there was a tendency for the ranking methods to produce less pronounced differences than the pairwise comparisons. However, the scale values derived from the three response collection methods agreed well for the castanets and the pop music. The accuracy of the evaluation of the perceived quality of the compression settings, therefore, does not differ much with respect to the methods employed.

With regard to the required time there were some differences, however. Pairwise comparisons took on average longer than the two ranking methods, but for the latter the variance between subjects was considerable: There were subjects who took more than twice as long as any subject with pairwise comparisons. For the castanets, the ranking by elimination method was significantly faster than ranking. Although showing the same pattern, this difference was not significant for the other two excerpts.

There were large individual differences in how often subjects played back the sounds. For castanets there was a tendency that sound buttons were played back less often with ranking by elimination than with ranking. The results were not significant, nor were they for the other two excerpts. The only exception might be the 64-kbps MP3 which seems to be especially easy to identify as the one with the lowest quality. One could therefore assume that the number of button clicks is correlated with the perceived difficulty of evaluating a given sound. The results for the time required suggest that ranking by elimination has a time advantage compared with ranking if stimuli are relatively easy to distinguish. The number of button clicks might also be considered as a measure of uncertainty. Participants replay those compression settings more frequently about which they are unsure, resulting in more sound button clicks for versions of higher quality. This does, on the other hand, not seem to improve the discrimination of these sound. Consequently, one may question whether the opportunity of repeated play back pays off.

As stated in the introduction, auditory attributes are latent psychological variables, and therefore in principle unobservable quantities. Scaling can thus only result from application of a psychological model. With paired-comparison data these models are usually easy to apply and interpret. Models for ranking data on the other hand require a large amount of observations in order for the distributional assumptions of a goodness-of-fit test to hold. Even with more than 50 subjects and only seven different stimuli it cannot be tested if the MBT model is adequate to describe the data. The approach used in this paper, namely justifying the MBT model by acknowledging that the BTL model was adequate to describe the paired-comparison data, is usually not applicable.

One of the reported findings suggests that scale

values for paired-comparison data differ more pronouncedly than for rankings. It is not quite clear why this result occurs. It could be that ranking procedures are in general less sensitive than paired comparisons. With paired comparisons, subjects are forced to evaluate all possible stimulus pairings. This results in a more elaborated information gathering, since subjects on average listen to the stimuli more often and compare all of them with each other. The second reason might be that with only a few observed rankings, the estimates of the MBT model might shrink towards the mean. Note that there are 5040 possible rankings with seven stimuli and that even with 52 subjects only a fraction of the possible rankings can be observed.

Summing up, it is plain that ranking methods are faster than paired comparisons. They seem not to lack accuracy; with sample sizes usually encountered in listening tests, however, the validity of ranking models can hardly be tested. Also, this speed comes at the price of large individual differences. Hence, it is difficult to predict how much time an individual participant will actually require. With paired comparison, although taking more time on average, one can exactly define the amount of time per subject. It has been shown that all three methods for sound quality evaluation derive similar result patterns concerning perceived quality of the stimuli but differ with respect to the time needed to gain these results. The herein introduced ranking-by-elimination procedure seems to be faster than a common ranking procedure when easy stimulus material is used. In order to corroborate these results the method should be applied and evaluated with a bigger set of stimuli that are relatively easy to discriminate.

5. REFERENCES

[1] Bradley, R. A. & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, *39*, 324–345.

[2] Choisel, S. & Wickelmaier, F. (2007). Evaluation of multichannel reproduced sound: scaling auditory attributes underlying listener preference. *Journal of the Acoustical Society of America*, *121*, 388–400.

[3] Critchlow, D. E. & Fligner, M. A. (1991). Paired comparison, triple comparison, and ranking experiments as generalized linear models, and their implementation on GLIM. *Psychometrika*, *56*, 517–533.

[4] David, H. A. (1988). *The method of paired comparisons*. New York: Oxford University Press.

[5] ITU-R BS.1116-1 (1994). Methods for the subjective assessment of small impairment in audio systems including multichannel sound systems. International Telecommunications Union, Geneva, Switzerland.

[6] ITU-R BS.1534-1 (2003). Method for the subjective assessment of intermediate quality level of coding systems. International Telecommunications Union, Geneva, Switzerland.

[7] ITU-T P.800 (1996). Methods for subjective determination of transmission quality. International Telecommunications Union, Geneva, Switzerland.

[8] Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York: Wiley.

[9] Mallows, C. L. (1957). Non-null ranking models: I. *Biometrika*, *44*, 114–130.

[10] Tversky, A. (1972). Elimination by aspects: a theory of choice. *Psychological Review*, *79*, 281–299.

[11] Tversky, A. & Sattath, S. (1979). Preference trees. *Psychological Review*, *86*, 542–573.

[12] Wickelmaier, F. (2008). Elimination-by-aspects (EBA) models. R package manual, version 1.5-2.

[13] Wickelmaier, F. & Schmid, C. (2004). A Matlab function to estimate choice model parameters from paired-comparison data. *Behavior Research Methods, Instruments, & Computers*, *36*, 29–40.

[14] Zieliński, S. K., Rumsey, F., & Bech, S. (2008). On some biases encountered in modern audio quality listening tests – a review. *Journal of the Audio Engineering Society*, *56*(6), 427–451.