Designing a system for Online Orchestra:

Microphone evaluation and cost-benefit analysis

Erik Geelhoed | David Prior | Michael Rofe

Abstract

Online Orchestra is a telematic performance project that aimed to enable young and amateur musicians to make music online. Part of the aim of the project was to assess the extent to which specialist equipment is needed in order to enable a high-quality musical experience in a telematic environment. This article reports a microphone evaluation study in which untrained participants were asked to assess ten characteristics of quality in five samples, each recorded using different microphone configurations. Results show that participants rated DPA VO4099 clip-on microphones best, but that a stereo pair of Sure SM57s – which are over ten times cheaper – was rated a close second. Multidimensional scaling also shows that SM57s had the highest similarity to all other microphone configurations, suggesting they are a good alternative to costlier solutions.

Keywords

Online Orchestra; microphone evaluation; cost-benefit analysis; telematic performance; perceived differences; audio characteristics

Introduction

Online Orchestra asked how burgeoning network technologies and creative approaches to composition could be used to enable young and amateur musicians living in remote locations access to large-scale ensemble music-making opportunities. The project resulted in a pilot performance involving four locations around Cornwall, United Kingdom. In the two preceding articles in this special issue of the *Journal of Music, Technology and Education* (see Prior et al. 2017a, 2017b), the computing and peripheral equipment deployed in the pilot performance is described, along with a rationale for these choices. The final design required many components – microphones, mixers, audio interfaces, speakers, cameras, screens, computers, routers, converters, wires/connectors and software – giving rise to many potential configurations.

A reasonable starting hypothesis is that higher cost equipment will deliver a higher quality audio-visual experience. Given that one of the aims of the project was to enable, as far as was possible, an *immersive* musical experience (see Rofe et al. 2017), audio-visual quality is notionally vital. However, reliance on high-cost equipment in turn contradicts a second aim of the project, which was to design a solution that was scalable, meaning a preference for low-cost equipment, or, better, equipment that future users might already own.

This article reports an experiment designed to investigate the cost-benefit of equipment that might be used in telematic performance, in order to gauge the necessity of higher cost technologies in the enablement of a high-quality musical experience. Time and budgetary constraints meant that it would not be possible to test an exhaustive range of equipment. It was therefore decided to focus on microphones, with a view that this study would form part of the decision-making process in Online Orchestra, but also that it might point more generally to the feasibility of telematic performance using lower cost equipment in the future. As such, this study aims to measure the extent to which untrained listeners could hear quality differences between microphones of highly differing cost. It should be noted from the outset that the objective was not to establish a single, 'ideal' microphone type or brand, but rather to develop an understanding of some guiding principles to take forward into the next phase of the project's development.

Microphone evaluations broadly fall into three categories. First, there are tests carried out by equipment manufacturers themselves, as well as independent assessments carried out by third parties to validate the claims made by these manufacturers. These evaluations vary in their methodological rigour but are invariably designed to produce quantitative data on the relevant characteristics of a product (e.g. audio bandwidth, polar response patterns or total harmonic distortion).[1] Second, and by far the most numerous and visible, are the reviews carried out by journalists and users, usually at the point of market release.[2] Occasionally, these will re-test the claims made by manufacturers, but, often, technical specifications will be quoted and assumed to be correct, with the review itself focused much more on the individual, subjective evaluation of the reviewer. Finally, there are studies carried out in the pursuit of academic research, whether they be in the fields of audio engineering,[3] acoustics or the social sciences. In the former two categories of academic research, evaluations might be based on emerging technology and innovation, where the latter tend to be less focused on individual items of equipment and more on higher level perceptual functioning of human subjects. The present study makes no attempt at the first of these three but integrates aspects of the second two.

Evaluations designed to assess the perception of audio quality often use the prevalent mean opinion score (MOS) method, ITU standard P800 (Kendrick et al. 2015; Carlile 1996). Although MOS is still the industry standard for measuring subjective audio quality assessment, these scales date back to a time before electronic calculators and computers and were thus designed for convenience. For some researchers, the MOS method now seems unnecessarily confining, with the Graphic Rating Scale (GRS) offering a more flexible alternative (Hayes and Patterson 1921). Originally, GRS questions were answered by asking participants to mark on a continuous line between two extremes (anchors). The GRS in this simplest of forms is also referred to as the Visual Analogue Scale (VAS). Subsequent developments of the GRS have involved partitioning the open scale into ten segments (Freyd 1923), eventually resembling the five-point Likert Scale (Likert 1932). The Unipolar Line Scale – closely resembling the original GRS and VAS – is now a standard of the International Standard Organization (ISO, 2003). GRSs are believed to counteract cognitive interference, asking participants to make a judgement based on an immediate response without having to think (Stone et al. 1974; Cloninger et al. 1976); they are simpler, map on better to underlying attitudes or emotions, are quicker to fill out and satisfy requirements for ratio-scale analysis (Anderson 1970). A line of about 100 millimetres in length (Geelhoed et al. 2000) can be helpful for analysis: tick marks are measured to one millimetre accuracy, the extreme left being 0 millimetres and the extreme right 100 millimetres, resulting in a 101-point ratio-scale.

As the rating scales lend themselves well to a ratio-scale analysis, the model of the normal or Gaussian distribution is employed. In the current study, participants were asked to give ratings on a scale of 0–100. One standard deviation (SD) can be expected to be around 25 (between 20 and 30). A rather unexpected finding in the current study was that although participants had never carried out a microphone assessment exercise of this kind before, the SDs for all questions were below 20, indicating an unusually high concordance amongst participants.

Method

Samples

A short (33 seconds) piece of music composed by Jim Aitchison for three violins, one clarinet and one alto saxophone were recorded simultaneously by five different microphone configurations and yielded the five samples shown in Figure 1.[4]

Figure 1: Microphone samples.

Sample	Short name	Details	Total cost
1	Shotgun	One *Rode NT3G* shotgun microphone (mono), placed approximately 4m back from performers, approximately 3m high	c£430
2	SM57s	Two *Sure SM57* cardioid dynamic microphones as coincident stereo pair, placed approximately 1.5m back from the performers, approximately 1.5m high	c£160

3	Decca Tree	Decca Tree comprising two *AKG c414 XLS* condenser microphones in cardioid mode as outriggers and a centrally mounted *Neumann u87* condenser also in cardioid mode, placed approximately 2m back from the performers, approximately 2.5m high	c£2800
4	Clip-ons	Five *DPA VO4099* clip-on hyper-cardioid condenser microphones, mounted directly onto the instruments and mixed to stereo	c£2115
5	SE4s	Two *SE Electronics SE4* condenser microphones with cardioid capsules mounted as coincident stereo pair. Microphones mounted above the Decca Tree (sample 3), approximately 2m back from the performers, approximately 3m high	c£450

The microphone configurations were chosen to reflect a variety of standard approaches to the capture of a small ensemble (the Decca Tree in sample 3; the individual 'spot' microphones in sample 4 and the coincident pair in sample 5), as well as two simple, low-cost 'wild-card' options (the shotgun microphone in sample 1 and the coincident dynamic microphones in sample 2), which were not anticipated to produce high-quality results. All microphones were recorded via *Digidesign DigiPre* microphone pre-amplifiers into *Digidesign 192* analogue-to-digital converters. No equalization or dynamic processing was used, but recordings were balanced to ensure equal RMS levels for each sample, within 1dB. Samples 3 and 4 required mixing of three and five microphones, respectively, down to stereo, so in these cases, subjective judgements were made in arriving at the final balance.

Participants

A total of 36 participants, three female and 33 male, took part, all music students at Falmouth University, taking the test as part of their class time. Twenty of the participants were 19 years of age (mean age = 19.97, SD = 3.975). Given that the aim of the test was to establish the extent to which future users of Online Orchestra (young and amateur musicians) might notice quality differences in microphones of different cost, all participants in the test were non-expert listeners. All were studying music technology so had some experience and knowledge of microphones, but all were at the early stages of their education, so their knowledge remained basic. Participants were not introduced to the aims and objectives of the Online Orchestra project in general, nor the aims of this test in particular, giving rise to a blind test. Participants undertook the experiment in two groups, 24 in group 1 and twelve in group 2; this variation in group size resulted from the numbers of students in classes.

Procedure

Group 1 assessed the samples in order 1: sample 1, sample 2, sample 3, sample 4 and sample 5. Group 2 assessed in reverse, order 2: samples 5, 4, 3, 2, 1. Participants sat in chairs in an acoustically treated recording studio, and the samples were played back as stereo recordings from Cockos Reaper via an *RME ADI-8* digital-to-analogue converter, through a pair of *Neumann KH310a* speakers, mounted on substantial speaker stands, placed approximately 3m apart. Participants could not see the computer screen displaying the software that presented the samples. Immediately after hearing a particular sample, participants were asked to rate aspects of its quality.

Questionnaire

Questions were presented in GRS format, taking care that phrasing was as simple and unambiguous as possible. For example:

> How easy was it to tell the instruments apart from one another?
>
> Not at all   Very
>
> |____|

Participants were asked to make a mark between the two extremes (including the two extremes). Participants were asked ten questions (see Figure 2), seven of which referred to audio quality and three that were more concerned with audio telepresence issues (questions 5, 8 and 9).

Figure 2: Questions.

Question number	Question	Short form in analysis
1	How easy was it to tell the instruments apart from one another?	Tell apart
2	Based on your knowledge of these instruments, how well do you think this recording represents the high frequencies?	High frequencies
3	Based on your knowledge of these instruments, how well do you think this recording represents the low frequencies?	Low frequencies
4	Based on your knowledge of these instruments, how ‘realistic’ did they sound overall?	Realistic
5	How close (physically) to the musicians did you feel?	Closeness
6	Were the instruments in this recording well balanced against each other?	Balance
7	How well did the instruments blend with each other?	Blending
8	How easy was it to tell what kind of room this recording	Kind of room

	was made in?	
9	How effectively did the recording enable you to feel like you were present in the room?	Presence
10	Based on a crude scale of ‘lo-fi’ to ‘hi-fi’, how would you rate this recording overall?	Quality

Statistical analysis

In the statistical analysis of differences, various forms of the analysis of variance (ANOVA) are used. The result tables provide the F-ratio, the accompanying degrees of freedom (DF) and probability (*p*-) values as well as Cohen’s (1973) partial eta-squared (η_p^2), which is a useful measure of the power/strength of the results. For the analysis of similarities, Pierson’s product moment correlation is used, denoted by the letter *r*, accompanied by ($N-1$) DF and *p*-value. Analysing a combination of correlation matrices to describe similarities between microphones, multidimensional scaling (MDS) cluster analysis is used (Young and Hamer 1987).

Results

Order effects

There were remarkably few order effects, and where these were found they were in unexpected places. Usually, order effects are prevalent in the ratings of the first and last samples: those who took part in order 1 would have to make an absolute judgement for sample 1 and, conversely, those in order 2 would have to do so for sample 5. After the initial sample in the given order, following judgements are expected to be relative.

Interestingly, three significant order effects were found for sample 2 (SM57s) and the participants in order 2 gave significantly higher ratings for (1) 'kind of room', (2) 'presence' and (3) 'quality' (see Figure 3). Thus in spite of the fact that this recording used the cheaper SM57s, was the fourth to be assessed by order 2 participants and was preceded by recordings using more expensive microphones (Decca Tree and Clip-ons), their judgement towards it was more favourable than the participants in order 1.

Figure 3: Order effects.

		Order	Effect	Order 1 ($N = 24$)		Order 2 ($N = 12$)	
Sample	Question	$F_{(1,35)}$	p	Mean	SD	Mean	SD
2: SM57s	Kind of room	4.553	0.040	40.0298	19.74786	54.0179	15.72729
2: SM57s	Presence	4.456	0.042	48.7723	19.25941	61.6071	11.77144
2: SM57s	Quality	4.345	0.045	49.9628	20.86553	63.3185	10.23310
4: Clip-ons	High frequencies	6.253	0.017	73.9211	10.36564	61.6071	19.36417
5: SE4s	Quality	9.261	0.004	47.2470	18.15725	65.3274	13.54531

Microphone sample characteristics

For each question, a 'within-subjects' ANOVA comparison was carried out between microphone configurations. Figure 4 shows the results of the microphone evaluation in order of high to low means across the five microphone ratings. Participants gave the highest ratings for 'high frequencies' and the lowest for 'kind of room'. The first row of the table shows that for the variable 'high frequencies', the overall difference between the samples (the main effect for microphone assessment) was highly significant: $F_{(4,136)} = 3.072$, $p = 0.019$, $\eta_p^2 = 0.083$; in the first column, the DF are shown, then the F-ratio, the level of significance (p), the strength of the results via partial eta-squared (η_p^2) and the overall mean, the average for all five microphone ratings. The questions probing aspects of telepresence are in italics. With the exception of ‘closeness’, there were for each question significant differences between microphones, shown in the column ‘p’ in bold, meaning significant main effects for type of microphone configuration.

Figure 4: Questions ordered by means of ratings.

Question	DF	F	*p*	η_p^2	Mean
High frequencies	4, 136	3.072	**0.019**	0.083	63.42348
Closeness	4, 140	1.754	0.142	0.048	61.41369
Realistic	4, 140	2.645	**0.036**	0.07	60.59028
Tell apart	4, 136	3.508	**0.009**	0.094	59.43878
Balance	4, 140	7.045	**0.000**	0.168	52.90675
Blending	4, 140	2.922	**0.023**	0.077	52.70337
Quality	4, 140	4.499	**0.002**	0.114	52.49008
Low frequencies	4, 136	2.436	**0.05**	0.067	51.72959
Presence	4, 140	4.894	**0.001**	0.123	50.90278
Kind of room	4, 140	3.298	**0.013**	0.086	43.68552

It is also revealing to focus on the discriminatory power of the participants (their ability to discern between microphone configurations), rather than on the magnitude of their ratings; Figure 5 ranks variables in order of magnitude of significance. Eta-squared (η_p^2) is a useful statistic for this purpose, although the *p*-values follow the same order.

Figure 5: Discriminatory power of participants.

Question	*p*	η_p^2
Balance	**0.000**	0.168
Presence	**0.001**	0.123
Quality	**0.002**	0.114
Tell apart	**0.009**	0.094
Kind of room	**0.013**	0.086
High frequencies	**0.019**	0.083
Blending	**0.023**	0.077
Realistic	**0.036**	0.07

Low frequencies	**0.05**	0.067
Closeness	0.142	0.048

Thus the participants were best at discerning which microphones were good at reflecting the balance between instruments than any other variable. The Clip-ons were rated best, which may sound as common sense given their proximity to instruments, but, surprisingly, the SM57s were a good second. (Tele)Presence also resulted in clear distinctions between samples, with the same two microphones (Clip-ons, SM57s) rated highest.

Follow-up analysis

For each question, a follow-up analysis was carried out to measure the degree of perceived difference between samples. As examples, results for 'high frequencies' and 'closeness' are discussed in detail. For 'high frequencies', the overall difference between samples (the main effect for microphone assessment) was highly significant: $F_{(4,136)} = 3.072$, $p = 0.019$, $\eta_p^2 = 0.083$. As shown in Figure 6, the highest ratings for how well the microphones represented the high frequencies were given to the Clip-ons (70.13), followed by the SM57s (66.17). The SE4s, Shotgun and Decca Tree resulted in ratings around 60 on a scale where 0 = 'not at all' and 100 = 'very'. The SDs (in the column SD) were surprisingly low, indicating a high level of concordance amongst the participants. This question resulted in a mean overall rating of 63.42.

Figure 6: High frequencies ratings.

Microphone	Mean	SD
Clip-ons	70.1276	15.03632
SM57s	66.1735	16.77634
SE4s	61.0459	16.0628
Shotgun	60.1786	17.7679
Decca Tree	59.5918	16.88302
Mean total	63.42348	

Figure 7 shows the mean ratings for the five microphone configurations, where (on the *Y*-axis) 0 = 'not at all' and 100 = 'very'.

Figure 7: High frequencies.

A paired comparison follow-up analysis was also undertaken, using a paired sample *t*-test; Figure 8 shows *p*-values (significant *p*-values in bold). Guided by this table, three bands of ratings were distinguished:

1. high: Clip-ons (dark tone in graphs)
2. mid: SM57s (mid tone in graphs)
3. low: the group of remaining three, SE4s, Shotgun and Decca Tree (light tone in graphs).

Thus, the Clip-ons' ratings are significantly higher than the SE4s, Shotgun and Decca Tree; there are no significant differences between the SE4s, Shotgun and Decca Tree. The SM57s take up an intermediary position; close inspection of Figure 8 reveals that the distinction between the Clip-ons and the SM57s is not clear cut. In a similar vein, the SM57s ratings are significantly different from the Shotgun ratings, but not from the SE4s and Decca Tree ratings.

Figure 8: Paired comparisons for high frequencies.

	Clip-ons	SM57s	SE4s	Shotgun
SM57s	0.313			
SE4s	**0.009**	0.179		
Shotgun	**0.007**	**0.044**	0.834	
Decca Tree	**0.004**	0.134	0.61	0.88

The next highest ratings (overall mean = 61.41) were given for 'closeness'. Repeating the above procedure gives rise to the results in Figure 9 and Figure 10. Again, SDs are narrow, indicating high concordance amongst the participants.

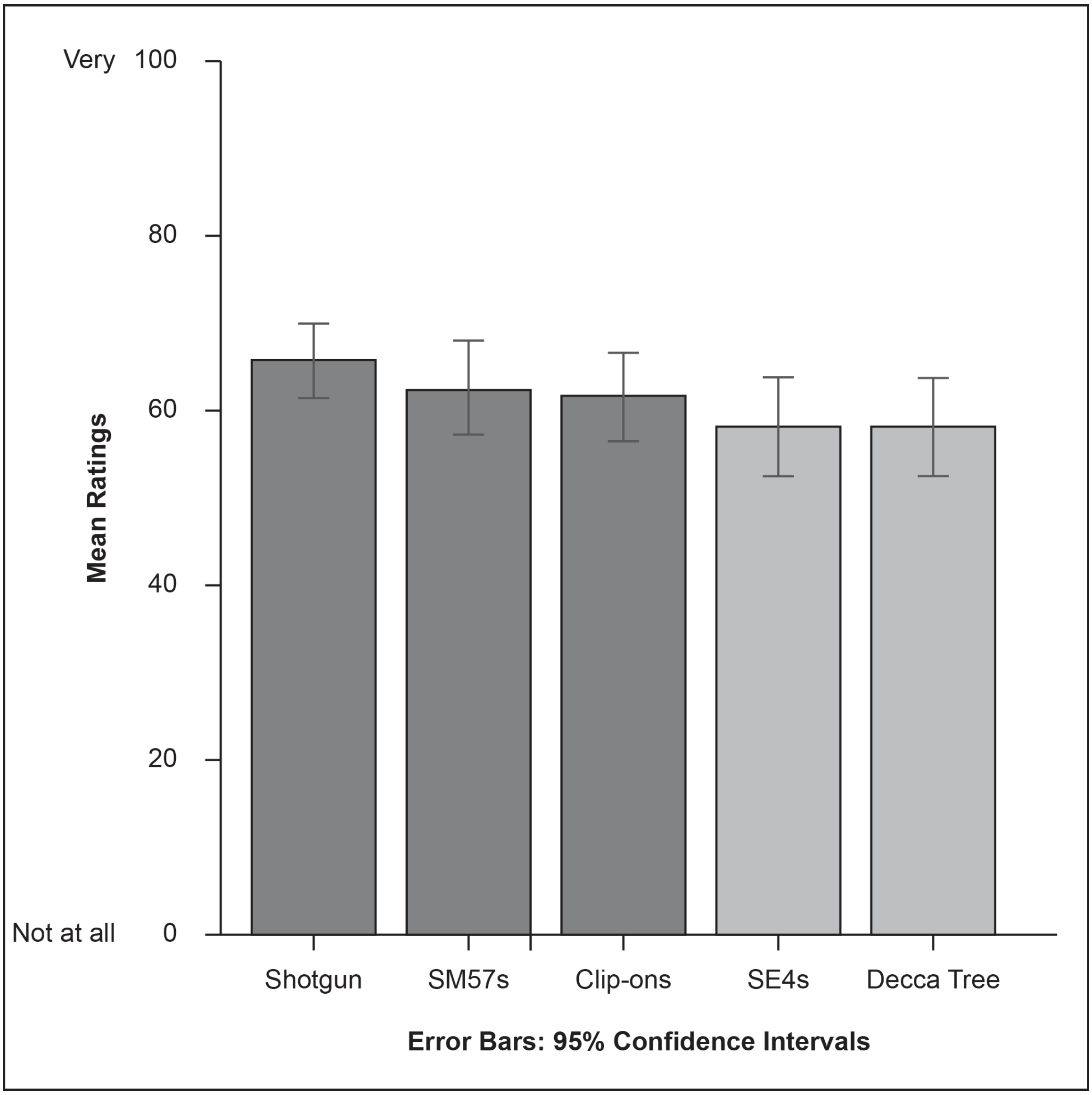
Figure 9: Closeness ratings.

Microphone	Mean	SD
Shotgun	65.9226	12.56996
SM57s	62.7232	15.99098
Clip-ons	61.8304	14.91041
SE4s	58.4077	17.06986
Decca Tree	58.1845	17.33615
Mean total	61.41369	

It is interesting that this question, probing telepresence, received high ratings, although the overall effect was not significant: $F_{(4,140)} = 1.754$, $p = 0.142$, $\eta_p^2 = 0.048$. Based on the paired comparison exercise, it is still possible to distinguish two bands of ratings:

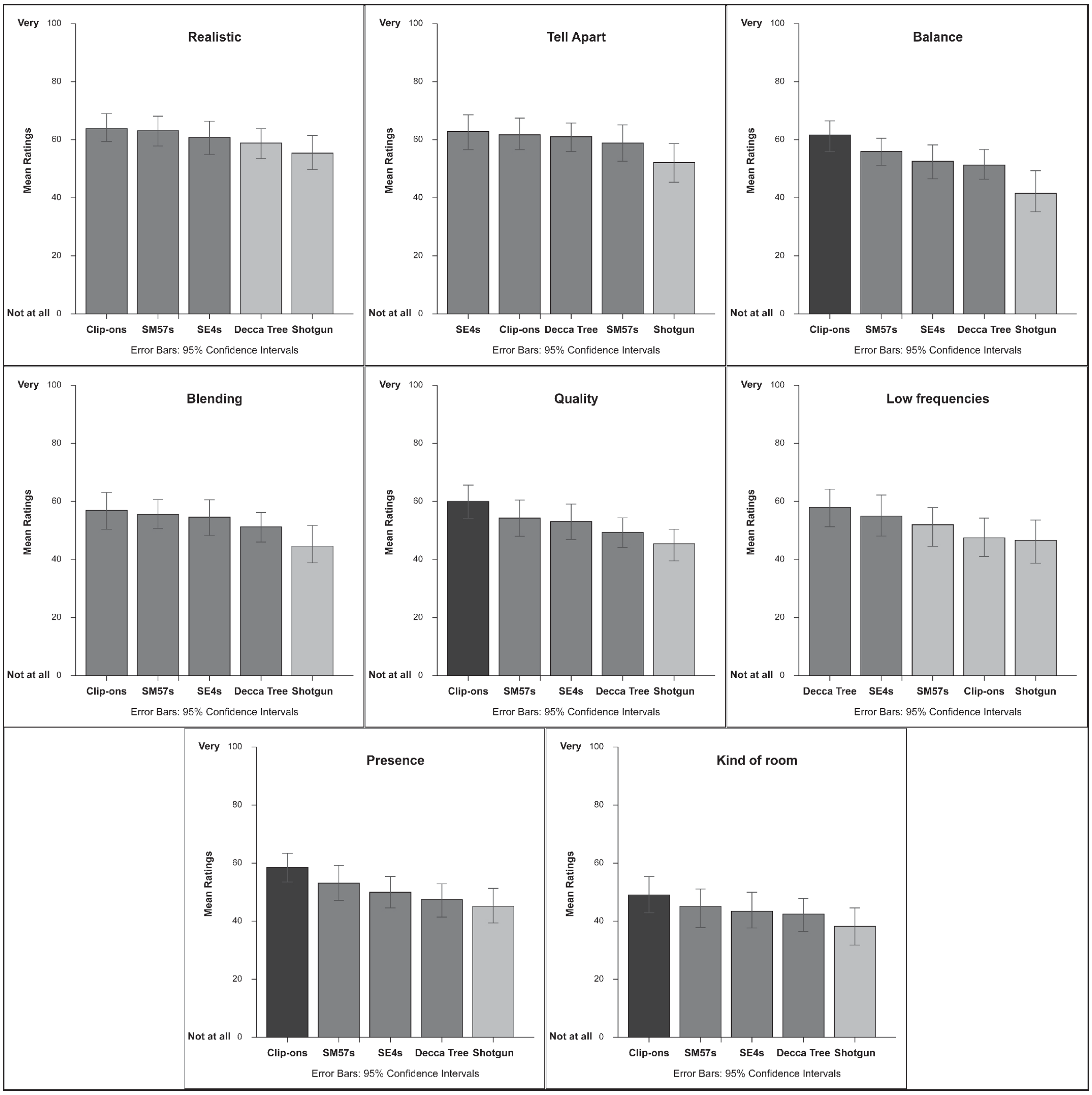
1. mid: Shotgun, SM57s and Clip-ons (mid tone)
2. low: SE4s, Decca Tree (light tone).

Figure 10: Closeness.



Repeating this procedure, Figure 11 shows the resulting graphs for the remaining eight questions.

Figure 11: Remaining characteristics.



Guided by the follow-up analysis, tied ranks (corresponding to shading in the bar charts) were assigned to the ratings. Figure 12 shows very clearly the superiority of the Clip-ons, as this was the only microphone configuration ranked as number 1 (‘high frequencies’, ‘balance’, ‘quality’, ‘presence’ and ‘kind of room’). Again, the SM57s fall second.

Figure 12: Tied ranks.

Questions	Clip-ons	SM57s	SE4s	Decca	Shotgun
High frequencies	**1**	2	4	4	4
Closeness	2	2	4.5	4.5	2
Realistic	2	2	2	4.5	4.5
Tell apart	2.5	2.5	2.5	2.5	5
Balance	**1**	3	3	3	5
Blending	2.5	2.5	2.5	2.5	5
Quality	**1**	3	3	3	5
Low frequencies	1.5	1.5	4	4	4
Presence	**1**	3	3	3	5
Kind of room	**1**	3	3	3	5
Totals	15.5	24.5	31.5	34	44.5
Total rank	**1**	**2**	**3**	**4**	**5**

Correlations between microphone configurations per question

This section explores whether there are similarities (or not) between microphone configurations, as measured by correlations in the way participants judged the samples for each question. For instance, question 8 (‘How easy was it to tell what kind of room this recording was made in?’) resulted in the highest *number* of significant correlations, and these correlations were also *highly* significant. Given that there were no missing data (the correlations resulted from ratings by 36 participants; the DF were N−1 = 36−1 = 35), Figure 13 reveals a high number of significant correlations, all positive.

Figure 13: Correlation table – kind of room.

		Shotgun	SM57s	Decca Tree	Clip-ons
SM57s	*r*	**0.592**	**1**		
	p	**0.000**			
Decca Tree	*r*	**0.571**	**0.647**	**1**	
	p	**0.000**	**0.000**		

Clip-ons	*r*	**0.303**	**0.585**	**0.519**	**1**
	p	**0.072**	**0.000**	**0.001**	
SE4s	*r*	**0.394**	**0.466**	**0.498**	**0.449**
	p	**0.018**	**0.004**	**0.002**	**0.006**

For instance, the ratings for this question between the Shotgun and the SM57s resulted in $r_{(\text{df } 35)} = 0.592$, $p = 0.000$. Figure 14 plots responses for 'kind of room' for each participant for the SM57s (along the *X*-axis) against those for the Shotgun. To reiterate, 0 = 'not at all' and 100 = 'very'. Results in the bottom left thus show participants who rated 'kind of room' very low for both the SM57s *and* the Shotgun; participants in the top right rated 'kind of room' very high for both the SM57s *and* the Shotgun. It is clear that, for the majority of participants, there is a strong (although not necessarily causal) relationship between the two sets of responses: those who rate the SM57s low for 'kind of room' also rate the Shotgun low; conversely, those who rate the SM57s high do so likewise for the Shotgun. The line through the centre of the scatterplot illustrates this positive correlation.

The following arbitrary weight is assigned to each level of significance in Figure 13 above, giving rise to Figure 15:

- at 0.1%, *p*-values of 0.001 and lower, weight = 4
- at 1%, *p*-values between 0.001 and 0.01 ($0.001 < p = <0.01$), weight = 3
- at 5%, *p*-values between 0.01 and 0.05 ($0.01 < p = <0.05$), weight = 2
- at 10%, *p*-values between 0.05 and 0.10 ($.05 < p = <0.10$), weight = 1.

Figure 14: Scatterplot of SM57s vs Shotgun – kind of room.

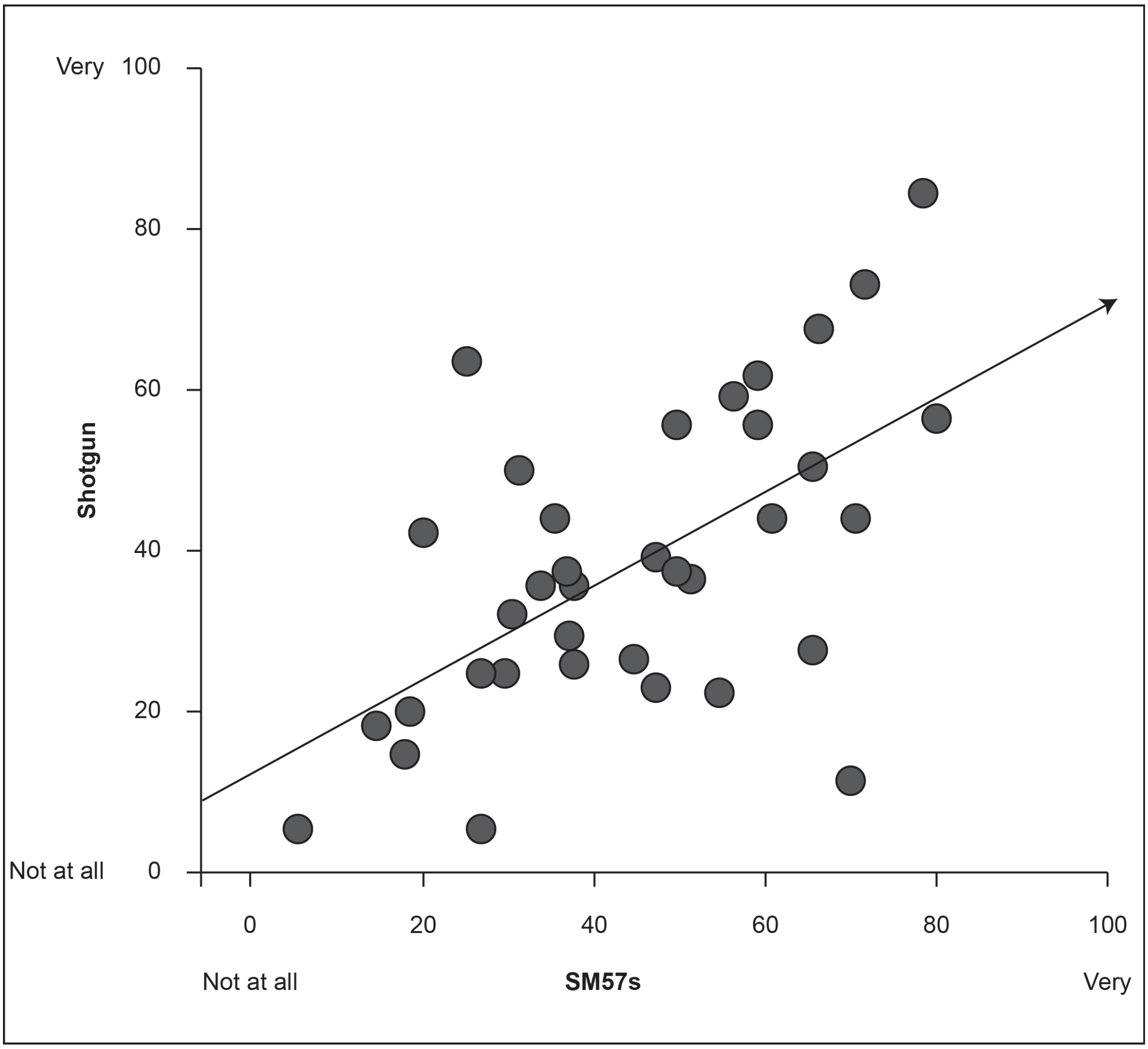


Figure 15: Weights assigned to significant *p*-values between microphones – kind of room.

	Shotgun	SM57s	Decca Tree	Clip-ons	SE4s	
SM57s	**4**					
Decca Tree	**4**	**4**				
Clip-ons	**1**	**4**	**3**			
SE4s	**2**	**3**	**3**	**3**		
Total weight	11	15	15	12	11	**64 (mean 6.4)**

For the Shotgun, there are two *p*-values significant at 0.1%: one at 5% and one at 10%, equalling a total weight of (2×4) + (1×2) + (1×1) = 11. Inspecting the column labelled 'Shotgun', a total weight of 11 can be seen in the bottom row. Following the entries for Clip-ons, *p*-values of 0.072 (again), 0.000, 0.001 and 0.006 can be seen, resulting in a weight of 12. In this way, each cell is counted twice. For the whole table, the average weight for the five microphones is calculated as 11 + 15 + 15 + 12 + 11 = 64/(5×2) = 6.4.

Figure 16 shows this mean total weight for each of the questions. Certain questions barely showed any significant correlations ('quality', 'low frequencies', 'high frequencies', 'balance'), and for 'blending', there were none.

Figure 16: Mean correlation weight per question.

Questions	Mean weight
Kind of room	*6.4*
Tell instruments apart	3.6
Realistic	3.4
Presence	*2.4*
Closeness	*1.4*
Quality	1
Low frequencies	0.9
Balanced	0.6
High frequencies	0.4
Blending	0

Interestingly, questions probing aspects of telepresence (shown in italics in Figure 16) resulted in more significant correlations than those purely interrogating auditory aspects of the samples. Adding up all the weight values for all the questions derives a similarity matrix between pairs of microphones, shown in Figure 17.

Figure 17: Total weights across all questions (similarity matrix).

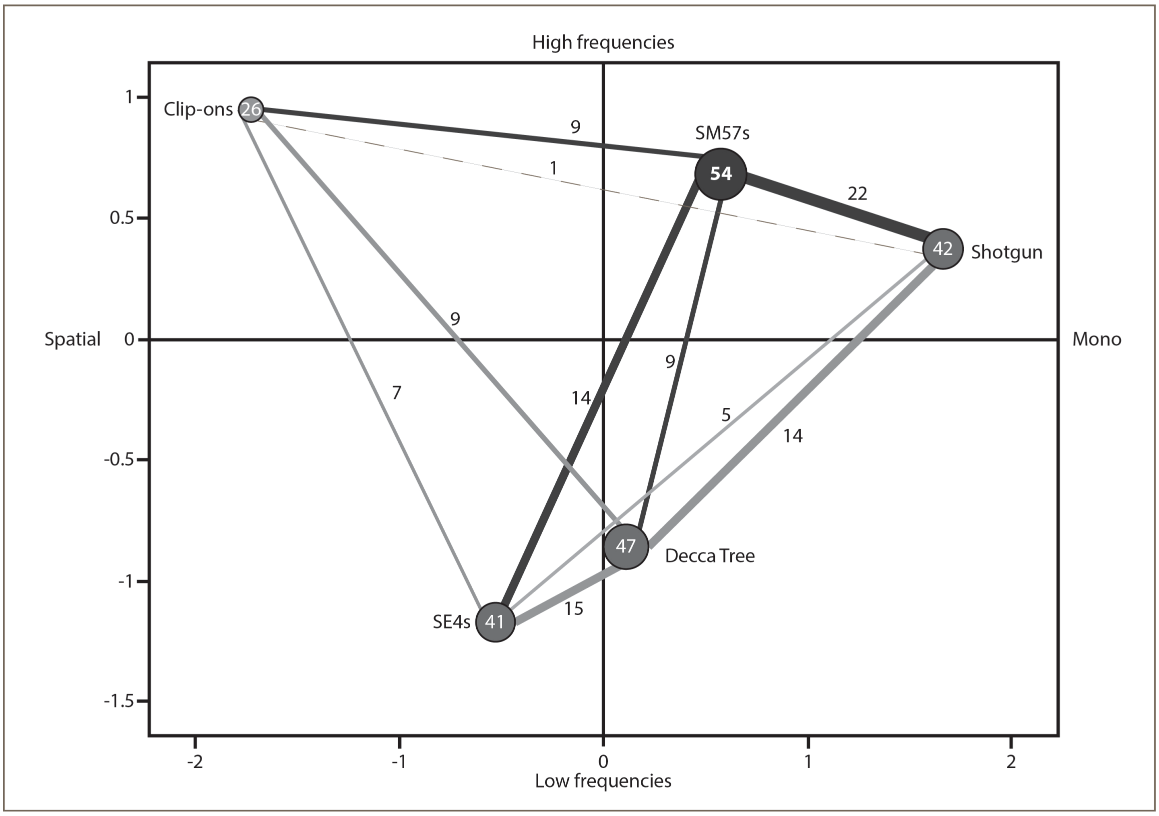
	Shotgun	SM57s	Decca Tree	Clip-ons	SE4s
Shotgun					
SM57s	22				
Decca Tree	14	9			
Clip-ons	1	9	9		
SE4s	5	14	15	7	
Total	42	54	47	26	41

Across all the questions, the Shotgun and SM57s are answered in the most similar way (22), and least similarity can be found between the Shotgun and the Clip-ons (1).

Using MDS (cluster analysis, Young and Hamer 1987), the data in Figure 17 can be plotted as a graph; see Figure 18. The sizes of the circles reflect the total weights. The SM57s (dark tone) have the largest circle (total weight = 54), and the Clip-ons (light tone) have the smallest (total weight = 26). Thus Figure 17 forms the basis for the MDS plot of Figure 18, visualizing the various distances between pairs of microphones in a two dimensional space.

The SM57s revealed most similarities with other microphones. Thus, in addition to being an overall good second (to the Clip-ons), the SM57s share most characteristics with the other microphones (weight = 54). The thickness of the lines from the SM57s to the other microphones represents the level of similarity, as shown in Figure 17. For instance, the line between the SM57s and Shotgun is the thickest, and the number 22 represents the cell of the intersection between SM57s and Shotgun. The plot also depicts how the Clip-ons are somewhat isolated from all other microphones; they are rated the highest (Figure 12) and also the least similar to others (Figure 18).

Figure 18: Visualizing similarities between microphones.



Shared characteristics between similar microphones enable patterns to be observed, leading to the labelling of axes in Figure 18. The extremes of the *X*-axis are occupied by the monophonic Shotgun microphone on the right and the more 'spatial' quality of individually mic'ed instruments using Clip-ons on the left. Considering that the Decca Tree and SE4s delivered a richer experience with regards to lower frequencies and that the clip-ons, SM57s and Shotgun received higher ratings for the high frequencies, the *Y*-axis has been labelled with extremes of low and high frequencies.

Discussion

In order to consider the cost-benefit of different microphones, 36 participants, all untrained listeners, were asked to judge ten quality-related characteristics of five different microphone configurations. Using a short musical sample, recorded simultaneously using five microphone configurations of different cost, the aim was to establish the extent to which participants perceived quality difference between recordings. Two high-cost solutions were tested: Clip-on microphones for each instrument (£2115), mixed down as a stereo recording, and a studio quality Decca Tree (£2800) suspended above musicians. Mid-cost solutions involved a Shotgun microphone (£430) and a pair of SE4 condenser microphones (£450). A pair of SM57 dynamic microphones (£160) constituted a low-cost solution.

Participants in the study judged the Clip-ons to be the best microphones in most respects, with the significantly cheaper SM57s coming a close second. Despite being the most expensive solution, the Decca Tree was only judged superior in one respect, this being a richer low frequency response, and was ranked third overall. In terms of specific characteristics, participants were best at discerning which microphones were good at reflecting the balance between instruments: again, the Clip-ons were rated best, with the cheaper SM57s a good second. Also of interest is that the Shotgun microphone was best at generating a feeling of closeness: an important aspect of telepresence. Evaluating similarities between microphones using cluster analysis, the cheaper SM57 revealed more similarities with other microphones than any others. That is to say, many characteristics of other microphones are reproduced well by the SM57s, leading to the conclusion that the SM57s are not only ranked second, but are also a *good all-round* configuration, modelling well the characteristics of more expensive alternatives.

The good performance of the Clip-ons confirmed expectation – placement directly on the instruments was likely to enable greater clarity, and this indeed resulted in higher scores across a number of characteristics. However, the good performance – and second place rank – of the significantly cheaper SM57s did come as a surprise. As such, it was necessary to ascertain how accurate and united participants were in this judgement. One of the first questions to arise concerns the ability of the participants to translate their auditory experience into a graphical rating. As auditory and visual processing resides in different areas of the brain, it would not be unreasonable to expect considerable variability amongst the participants' responses. In addition, it might be expected that participants would not be able to distinguish clearly between the various recordings of the piece of music: participants were non-specialist listeners (which is to say they had little or no previous experience of audio evaluation tests), chosen as a reflection of the type of user (young and amateur musicians) who would be expected to perform in Online Orchestra.

As such, two specific questions emerge: (1) how agreed are the participants amongst themselves? (2) how can their discriminatory power to distinguish between microphone configurations be assessed? If the participants revealed a wide variation in the way they judged the various aspects of the musical samples, then wide SDs (a measure of spread) in the distribution of responses would be expected. If that were to be the case then, using a scale ranging from 0 (not at all) to 100 (very), SDs in excess of 30 might be expected. However, for all 50 questions (five samples; ten questions per sample), SDs well below 20 were found, indicating an unusual concordance amongst the participants. In other words, participants displayed a high level of agreement in the way they answered questions.

It is not unreasonable to expect that discriminating between short samples of recorded (identical) music using the five different microphone configurations is a difficult task, particularly for non-specialist listeners, and might result mostly in non-significant statistical differences (even if the SDs were narrow). However, highly significant main effects were found for microphone configuration, with the exceptions of ratings for ‘low frequencies’ which was significant at 5 per cent and ‘closeness’ which was non-significant. The high number of significant differences that were found is all the more surprising given that participants tended on the whole not to assign extremely high or low ratings. That is to say, the means for each question were rather conservative, ranging from 40 to 60 (roughly); a narrower range has a lower likelihood to produce significant differences.

Conclusion

As discussed in Rofe et al. 2017, one of the aims of Online Orchestra was to develop a solution to telematic performance that could be scaled, enabling future users to take part without reliance on specialist equipment that might come at high cost. In this context, the finding that the SM57s performed a close second to the Clip-ons when assessed by untrained listeners is a striking result. In fact, taking account of all ten variables, the mean overall ratings for each microphone configuration are Clip-ons, 58.62; SM57s, 57.75;

SE4s, 56.13; Decca Tree, 54.06; and Shotgun, 49.53. So even though the Clip-ons are over ten times the cost of the SM57s, the difference between the overall mean ratings of these two configurations is negligible.

As described in the introduction to this article, the aim here is not to make recommendations regarding ideal microphones, but rather to assess the cost-benefit of different types of microphone. Moreover, this study has not tested microphones in a telematic context, in which technical characteristics such as feedback rejection might come more into play (see Prior et al. 2017a). However, in terms of judgements made by untrained listeners about sonic characteristics alone, this study suggests strongly that, although the more expensive Clip-ons were ranked the highest, the significantly cheaper SM57s offer a perfectly reasonable alternative were cost to be a consideration.

Time and budgetary constraints meant that the focus of this study has been confined to microphone evaluations. However, telematic performance relies equally on speakers, cameras and screens, not to mention additional processing brought about through computing. As such, this study is indicative of the type of approach that could be taken in future research to test the cost-benefit of other variables, with the aim overall of defining a minimum technical specification that is acceptable for young and amateur musicians. The present study offers promising potential in this regard, in its unexpected finding that the low-cost SM57 – a microphone that many schools and community groups might already in fact own – performs highly similarly to several higher cost alternatives.

References

Anderson, N. H. (1970), 'Functional measurement and psychophysical judgment', *Psychological Review*, 77:3, pp. 153–70.

Art of Record Production (2017), 'The 12th Art of Record Production Conference

Mono: Stereo: Multi' www.artofrecordproduction.com/index.php. Accessed 22 January 2017.

Audio Engineering Society (2017), www.aes.org. Accessed 22 January 2017.

Carlile, S. (1996), 'The physical and psychophysical bias of sound localization', in S. Carlile (ed.), *Virtual Auditory Space: Generation and Application*, Berlin: Springer-Verlag, pp. 27–78.

Cloninger, M.R., Baldwin, R.E., and Krause, G.F. (1976), 'Analysis of Sensory Rating Scales', *Journal of Food Science*, 41, pp. 1225-8.

Cohen, J. (1973), 'Eta-squared and partial eta-squared in fixed factor ANOVA designs', *Educational and Psychological Measurement*, 33:1, pp. 107–12.

Freyd, M. (1923), 'The graphic rating scale', *Journal of Educational Psychology*, 14:2, pp. 83–102.

GearSlutz (2017), www.gearslutz.com. Accessed 22 January 2017.

Geelhoed, E. N., Falahee, M. and Latham, K. (2000), 'Safety and comfort of eye glass displays', *Proceedings Second International Symposium, Handheld and Ubiquitous Computing*, Bristol, UK, 25-27 September, London: Springer, pp. 236–47.

Geelhoed, E. N., MacRae, A. W. and Ennis, D. M. (1993), ‘Preference gives more consistent judgments than oddity only if the task can be modeled as forced choice’, *Perception & Psychophysics*, 55:4, pp. 473–77.

Geelhoed, E. N., Parker, A., Williams, D. J. and Groen, M. (2009), ‘Effects of latency on telepresence’, HP labs technical report: HPL-2009-120, http://www.hpl.hp.com/techreports/2009/HPL-2009-120.html. Accessed 14 November 2016.

Geelhoed, E. N., Stenton, P., Singh-Barmi, K. and Biscoe, I. (2014), ‘Necessidades dos usuários de espaços de performances imersivas mediatizadas’ (‘User requirements in immersive mediated performance’), *Revista Mapa*, 1:1, pp. 96–119.

Hayes, M. H. and Patterson, D. G. (1921), ‘Experimental development of the graphic rating method’, *Psychological Bulletin*, 18, pp. 98–99.

HPL, ‘Effects of latency on telepresence’ http://www.hpl.hp.com/techreports/2009/HPL-2009-120.html. Accessed 14 November 2016.

Kendrick, P., Jackson, I. R., Fazenda, B. M., Cox, T. J. and Li, F. F. (2015), ‘Microphone handling noise: Measurements of perceptual threshold and effects on audio quality’, *PLoS One*, 10:10, e0140256. Accessed 14 November 2016.

Likert, R. (1932), ‘A technique for the measurement of attitudes’, *Archives of Psychology*, 140, pp. 1–55.

Microphonedata.com (2017), www.microphonedata.com. Accessed 22 January 2017.

Munshi, J. (1990), *A Method for Constructing Likert Scales*, California: Sonoma State University.

Prior, D., Reuben, F., Biscoe, I. and Rofe, M. (2017a), 'Designing a system for Online Orchestra: Computer hardware and software', *Journal of Music, Technology and Education*, 10: 2-3, pp. 185-96.

Prior, D., Reeder, P., Rofe, M., Biscoe, I. and Murray, S. (2017b), 'Designing a system for Online Orchestra: Peripheral equipment', *Journal of Music, Technology and Education*, 10: 2-3, pp. 197-212.

Rofe, M., Murray, S. and Parker, W. (2017), 'Online Orchestra: Connecting remote communities through music', *Journal of Music, Technology and Education*, 10: 2-3, pp. 147-66.

Stone, H., Sidel, J., Oliver, S., Woolsey, A. & Singleton, R.C. (1974), 'Sensory Evaluation by Quantitative Descriptive Analysis', *Food Technology,* Nov 1974, 24-34.

Sound On Sound (2017), http://www.soundonsound.com. Accessed 22 January 2017.

Young, F. W. and Hamer, R. M. (1987), *Multidimensional Scaling: History, Theory, and Applications*, London: Lawrence Erlbaum Associates.

Notes

[1] Almost every microphone to be released will feature tests of this kind. A comprehensive archive of microphone data can be found at microphonedata.com.

[2] A popular example of a UK-based magazine specializing in equipment reviews is *Sound on Sound*, which frequently reviews new product releases. Test data will often be quoted and occasionally new products will be subject to 'bench tests' as part of the review. However, the focus of these articles will always be subjective and applied to 'real

world' session contexts. Online forums, such as Gearslutz, also feature ongoing discussions about individual microphones. While new product releases will often stimulate discussion, threads are as likely to be initiated by a question about an established model and will generally tend towards subjective analysis.

3. As an academic discourse, the field of audio engineering is approached both from a technical perspective through organizations such as the Audio Engineering Society (AES 2017) and from an Arts and Humanities perspective by the Association for the Art of Record Production (ASARP 2017).

4. This combination of instruments was chosen (1) to model partially the final Online Orchestra ensemble and (2) to provide a range of timbres and dynamic levels.