AN EXPERIMENT ABOUT ESTIMATING THE NUMBER OF INSTRUMENTS IN POLYPHONIC MUSIC: A COMPARISON BETWEEN INTERNET AND LABORATORY RESULTS

Michael Schoeffler, Fabian-Robert Stöter, Harald Bayerlein, Bernd Edler, Jürgen Herre
International Audio Laboratories Erlangen
michael.schoeffler@audiolabs-erlangen.com

ABSTRACT

Internet experiments in the fields of music perception and music information retrieval are becoming more and more popular. However, not many Internet experiments are compared to laboratory experiments, the consequence being that the effect of the uncontrolled Internet environment on the results is unknown. In this paper the results of an Internet experiment with 1168 participants are compared to those of the same experiment with 62 participants but previously conducted in a controlled environment. The comparison of the Internet and laboratory results enabled us to make a point on whether the Internet can be used for our experiment procedure. The experiment aimed to investigate the listeners ability to correctly estimate the number of instruments being played back in a given excerpt of music. The participants listened to twelve short classical and pop music excerpts each composed using one to six instruments. For each music excerpt the participants were asked how many instruments they could hear and how certain they were about their estimation.

1. INTRODUCTION

In psychoacoustics, a sequence of sounds grouped by the auditory system is known as an "auditory stream" which was coined by Bregman and Campbell [2]. In the past decades, a lot of experimental work related to "auditory streams" has been done [1]. A majority of these experiments were psychoacoustically motivated, e. g. the stimuli used were mostly of simple type like sinusoids or noises. Especially from a psychoacoustic point of view, music is a very complex sound signal which contains high-level information (e. g. instrumentation and song lyrics). When listening to music, this high-level information is also mentally processed by humans. For developing auditory models, it could be helpful to know the maximum number of instruments humans are able to estimate when listening to music.

For many types of music perception experiments such as estimating the number of instruments, it is essential that the selected participants represent a large population. As recent research in cross-cultural music perception and cognition reveals: The perception of music is dependent on the origin of people [14]. Besides the cultural background, other aspects like their profession might have an influence when estimating the number of instruments being played back, e. g. musicians might recognize instruments much more easily since they are in touch with instruments in their everyday life. One of the advantages of Internet experiments (also called web-based experiments or web experiments) is that it is easier to gather participants with different backgrounds and from different regions than in laboratory experiments. For a summary of benefits and disadvantages of Internet experiments see [11]. In music perception a major argument against Internet experiments is that there is no control about the environment. With the spreading of mobile devices with Internet connections this argument becomes more apparent, since the environment of the participants can range from a quiet place at home to a noisy place outside.

By comparing the results of the Internet experiment presented in this paper to the results of the same experiment but previously conducted in a controlled environment [7], we contribute to answering the research question, whether the Internet can be used for experiments in music perception. Furthermore, subpopulations like headphones-users and loudspeaker-users are examined whether they lead to more reliable responses.

2. RELATED WORK

The ability of estimating the number of instruments is probably related to the ability of auditory stream segregation. An overview of auditory stream segregation in general is given by Bregman [1] and Wang and Brown [15].

In 1989, Huron conducted a musically motivated experiment related to stream segregation [9]. In his experiment he asked the participants for the number of voices in excerpts of organ music. Huron defined a voice as a single "line" of sound, more or less continuous, that maintains a separate identity in a sound field or musical texture (an overview of voice definitions is given by Cambouropoulos [3]). Huron came to the conclusion that the number of correctly identified voices is up to three. Based on Hurons work, we carried out an experiment where we asked musi-

cians and non-musicians how many instruments they could hear in short pieces of music [7]. In contrast to Huron we addressed a more general case where voices are played by different instruments and not only by organ. In Huron's experiment the responses of the participants were time-varying for each stimulus. Another difference is that in Huron's experiment all six participants except one had a musical background whereas our experiment was also attended by a large group of non-musicians.

It has been becoming more and more popular to use the Internet for experiments in music or auditory perception. One of the first auditory experiments were conducted by Welch and Krantz [16] in 1996. A web experiment related to MIREX tasks using *Amazon Mechanical Turk* has been conducted by Lee [10]. An experiment with a large attendance was done by Salganik *et al.* [13] in 2006. They had over 14.000 participants and examined the social influence on participants in an artificial music market. An overview of recent Internet experiments is given by Reips [12].

3. METHOD

3.1 Stimuli

For the stimuli generation, MIDI files of music pieces from the RWC database [8] were selected. The MIDI files were modified so that each file has a specific number of instruments being played back. Since the number of instruments being played back had to be as constant as possible, a so-called "instrumental stationarity" for each music piece was calculated. The "instrumental stationarity" is a measure that shows whether all instruments are played the whole time for a given start position and length. See [7] for the detailed equation of "instrumental stationarity" and more information about the stimuli generation. The RWC files were manually filtered a priori to exclude items dominated by lead instruments or singing voices. From the remaining items thirteen excerpts of MIDI files were selected with the help of the "instrumental stationarity" having one to six simultaneously playing instruments. Table 1 shows the selected excerpts including their instrumentation. The duration of each excerpt is around seven seconds. By cutting at note offsets we varied the lengths of the excerpts to make them semantically more meaningful. Six items (notated as C0**) belong to the classical Western music genre whereas the other items are of mixed genre. The MIDI excerpts were humanized and rendered by a sequencer software utilizing state-of-the-art commercial sampling products into WAV files. The rendered files were processed with convolutive reverb to match the original recordings. In informal listening tests the quality of the renderings was evaluated. In addition, participants did not give negative feedback about the artificialness of the items during the laboratory experiment. Additionally a loudness normalization was applied according to EBU-R128 [5]. To avoid spatial cues the files were downmixed to mono at 16 bit/44.1 kHz.

RWC ID	Start [s]	Dur. [s]	Instrumentation	$\sum$
J021	46.5	6.6	Piano, Contrabass (pizz.) and Trumpet	3
C001	0.0	9.0	Bassoon	1
G047	35.3	8.3	Violoncello	1
C016	0.9	7.6	Viola and Violoncello	2
G068	132.4	6.6	Violin and Flute	2
C018	240.4	5.4	French Horn, Piano and Violin	3
G046	0.3	7.9	Contrabass, Piano and Violoncello	3
C013	5.6	6.0	Flute, Viola, Violin and Violoncello	4
G036	0.0	6.5	Acoustic Guitar, Electric Bass, Piano and Violin	4
C012	112.0	6.0	Contrabass, Flute, Viola, Violin and Violoncello	5
G037	67.1	7.0	Acoustic Guitar, Contrabass (pizz.), Flute, Piano and Tenor Sax	5
C001	147.8	6.0	Bassoon, Clarinet, Contrabass, French Horn, Oboe and Violin	6
G028	17.5	6.5	Electric Bass, Electric Guitar, Flute, Piano, Trombone and Trumpet	6

Table 1. Selected items from the RWC Music Database [8]. Item *J021* was used as training item.

3.2 Participants

Participation in the experiment was done by visiting the experiment's website [1] . The experiment was promoted in mailing lists, forums, social networks and by personal invitations. Most of the forums and mailing lists were audio-related. No material incentive was given to participants. For motivating the participants a high score was added to the experiment.

In total 1310 website visitors attended the experiment. We identified 115 of them as participants who did the experiment more than once by using a browser fingerprinting method (for more details in browser fingerprinting, see [6]). In this case, only the first trial is used in the results analysis. Our browser fingerprinting method created a hash value by using the visitor's browser settings, e. g. screen resolution and installed plugins. In addition, we excluded 27 participants since they gave at least one non-serious response. We defined responses with zero instruments (25 participants) and responses with more than 12 instruments (two participants) as invalid. After the screening we had 1168 valid participants.

The participants were asked by a questionnaire whether they have a professional background in audio, play at least one instrument (including singing) and are familiar with listening tests. Detailed information about the participants are described in Table 2.

3.3 Materials and Apparatus

The main functionality of the experiment's website was written in HTML5 and JavaScript. The website was tested for all major web browsers and optimized for mobile devices and desktop computers. The default file format for the stimuli was WAV. Since some browsers (e. g. Internet Explorer) did not support WAV, MP3 (encoded with 256 kbits/s CBR with Fraunhofer Encoder) was used as alternative file format. The alternative file format was only used when WAV was not supported by the browser.

[1] `http://www.audiolabs-erlangen.com/experiments/wice/`

Total	Age group	Musician	Professional	Headphones
1168	0 [0-12]	-	-	-
	110 [13-19]	74 [yes]	12 [yes]	5 [yes]
				7 [no]
			62 [no]	30 [yes]
				32 [no]
		36 [no]	0 [yes]	-
				-
			36 [no]	17 [yes]
				19 [no]
	889 [20-39]	463 [yes]	128 [yes]	81 [yes]
				47 [no]
			335 [no]	154 [yes]
				181 [no]
		426 [no]	46 [yes]	31 [yes]
				15 [no]
			380 [no]	164 [yes]
				216 [no]
	143 [40-59]	98 [yes]	32 [yes]	22 [yes]
				10 [no]
			66 [no]	31 [yes]
				35 [no]
		45 [no]	8 [yes]	5 [yes]
				3 [no]
			37 [no]	17 [yes]
				20 [no]
	26 [60+]	13 [yes]	6 [yes]	3 [yes]
				3 [no]
			7 [no]	4 [yes]
				3 [no]
		13 [no]	2 [yes]	2 [yes]
				0 [no]
			11 [no]	5 [yes]
				6 [no]

Table 2. Information about the participants.

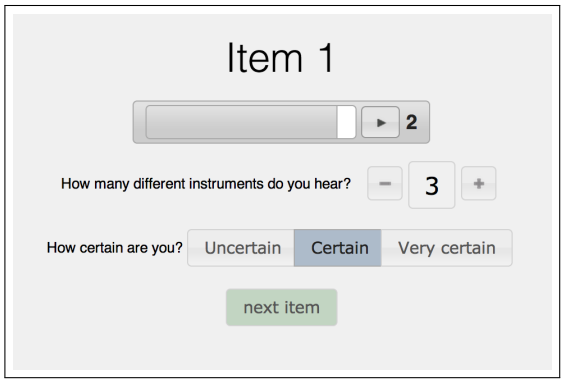

Figure 1. Experiment User Interface.

	Num_{Inst}						
Resp	$I=1$	$I=2$	$I=3$	$I=4$	$I=5$	$I=6$	n
$R=1$	2025	373	5	18	13	12	2446
$R=2$	298	1642	810	736	451	382	4319
$R=3$	12	277	1343	1145	1093	1069	4939
$R=4$	1	43	158	386	645	680	1913
$R=5$	0	1	18	48	120	155	342
$R=6$	0	0	1	3	12	30	46
$R>6$	0	0	1	0	2	8	11
	14016 responses (1168 participants · 12 items)						
Probability of $Resp_{\mathrm{Correct}}$	0.87	0.70	0.57	0.17	0.05	0.01	

Table 3. Responses from the participants. The cells with a gray background represent correct estimations.

3.4 Procedure

The experiment started on February the 15th, 2013 and lasted until March the 15th, 2013.

At first, participants filled out a short questionnaire. They were asked which audio setup they are using, whether they regularly play any musical instruments or do singing, have a background in professional audio, are familiar with listening tests and which age group they belong to.

After filling out the questionnaire, the participants did a short training. The purpose of the training was to familiarize the participants with the user interface and to give them the option to adjust the volume. The training had one stimulus with three instruments being played. The instruments were piano, bass and trumpet. The participants were told on the training page how many and which instruments are played back. During the training it was possible to listen to the stimulus unlimited times.

Followed by the training, the participants had to estimate the number of instruments being played in twelve stimuli. The experiment question was "How many different instruments do you hear?". Participants could listen to each stimulus up to three times. In addition, they were asked how certain they were in their response. They could choose between "uncertain", "certain" and "very certain". The user interface is shown in Figure 1.

After the participants estimated the number of instruments for all twelve stimuli, they were given a score based on their performance. Besides their personal score, a percentile rank showed how each participant performed compared to all the other participants.

4. RESULTS

The independent variables are the number of instruments being played back (Num_{Inst}), whether a participant is musical (*Musical*), professional in audio (*Professional*) and which setup was used (*Setup*). A participant is defined as musical ($Musical = true$) when he or she is regularly playing an instrument (including singing). The same applies to being professional ($Professional = true$) which is set when the participant responded that he or she is a professional in audio. The responses for the setup used can either be headphones ($Setup = `headphones'$) or loudspeaker ($Setup = `loudspeaker'$). The dependent variable is the participant's estimation of the number of instruments being played back (*Resp*). A correct estimation is defined as

$$Resp_{\mathrm{Correct}} = \begin{cases} 0 & \text{if } Num_{\mathrm{Inst}} \neq Resp \\ 1 & \text{if } Num_{\mathrm{Inst}} = Resp \end{cases}. \qquad (1)$$

Table 3 shows the responses of the participants for all stimuli.

For testing hypotheses, a logistic regression model with the response variable $Resp_{\mathrm{Correct}}$ and the predictor vari-

Coefficient	Estimate	Std. Error	z-value	p-value	Average Marginal Effects
(Intercept)	1.5015	0.06836	21.963	$<$ 2e-16	0.1886
$Num_{\text{Inst}} = 2$	-1.0293	0.07651	-13.453	$<$ 2e-16	-0.1293
$Num_{\text{Inst}} = 3$	-1.6021	0.07463	-21.466	$<$ 2e-16	-0.2012
$Num_{\text{Inst}} = 4$	-3.5674	0.08380	-42.572	$<$ 2e-16	-0.4481
$Num_{\text{Inst}} = 5$	-4.8759	0.11290	-43.188	$<$ 2e-16	-0.6125
$Num_{\text{Inst}} = 6$	-6.3061	0.19428	-32.459	$<$ 2e-16	-0.7922
Musical $= true$	0.5266	0.04932	10.676	$<$ 2e-16	0.0661
Professional $= true$	0.3306	0.06234	5.303	1.14e-07	0.0415
Setup $=$ '$headphones$'	0.1071	0.04823	2.220	0.0264	0.0135

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 18816 on 14015 degrees of freedom
Residual deviance: 11036 on 14007 degrees of freedom
AIC: 11054
Number of Fisher Scoring iterations: 7
McFadden's Pseudo R-squared: 0.413

Table 4. Logit regression model for response variable $Resp_{\text{Correct}}$ calculated from the data obtained by the Internet experiment.

ables Num_{Inst}, *Musical*, *Professional* and *Setup* was calculated. The estimated coefficients, p-values and average marginal effects are shown in Table 4. Average marginal effects in the regression model describe the increase in probability for correctly estimating the number of instruments when the predictor variable is increased by one level. Compared to the other coefficients the average marginal effect of *Setup* $=$ '$headphones$' is very low. By using headphones instead of loudspeakers it is 1.35% more likely to estimate the number of instruments correctly.

As expected, participants who play an instrument or do singing (*Musical* $= true$) performed slightly better than non-musicians. According to the average marginal effect their chance of estimating the number of instruments correctly is 6.61% more likely for all stimuli. A similar increase for estimating the number correctly (4.15%) can be seen for participants being a professional in audio (*Professional* $= true$).

The average marginal effects of Num_{Inst} indicates up to which point humans are able to correctly estimate the number of instruments being played back. The average marginal effect of $Num_{\text{Inst}} = 2$ shows that it is 12.93% less likely to estimate correctly when listening to two instruments instead of one instrument. Furthermore, in case of three instruments being played back the probability of estimating the wrong number increases to 20.12%. The highly negative average marginal effect of -0.4481 for $Num_{\text{Inst}} = 4$ indicates that it is becoming very unlikely for humans to estimate the number of instruments correctly compared to estimating the number of one to three instruments.

For a detailed analysis of the differences between the Internet experiment and the laboratory experiment in a controlled environment, a second logit regression model was calculated. This logit regression model includes the data of the previous experiment which has responses of 62 participants. Besides Num_{Inst} an additional predictor variable

Coefficient	Estimate	Std. Error	z-value	p-value	Average Marginal Effects
(Intercept)	2.0226	0.11718	17.260	$<$ 2e-16	0.2590
$Num_{\text{Inst}} = 2$	-1.0135	0.07424	-13.652	$<$ 2e-16	-0.1298
$Num_{\text{Inst}} = 3$	-1.5914	0.07223	-22.033	$<$ 2e-16	-0.2038
$Num_{\text{Inst}} = 4$	-3.4786	0.08031	-43.316	$<$ 2e-16	-0.4454
$Num_{\text{Inst}} = 5$	-4.8058	0.10919	-44.015	$<$ 2e-16	-0.6154
$Num_{\text{Inst}} = 6$	-6.2481	0.19033	-32.827	$<$ 2e-16	-0.8000
Environment $=$ 'web'	-0.1435	0.10569	-1.358	0.174	-0.0184

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 19826 on 14759 degrees of freedom
Residual deviance: 11819 on 14753 degrees of freedom
AIC: 11833
Number of Fisher Scoring iterations: 7
McFadden's Pseudo R-squared: 0.404

Table 5. Logit regression model for response variable $Resp_{\text{Correct}}$ calculated from the data obtained by the Internet experiment and the laboratory experiment.

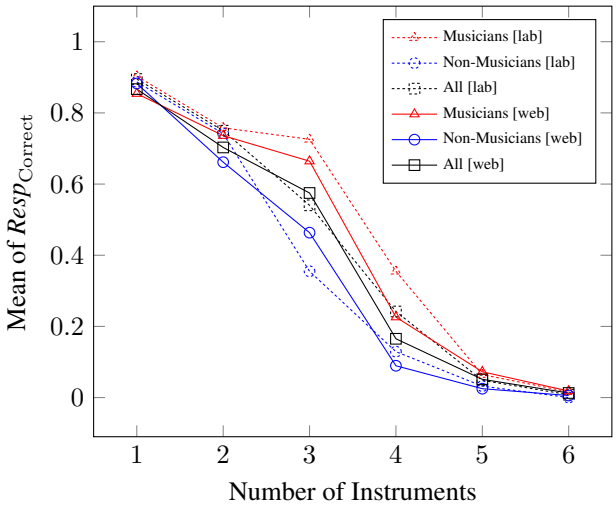

Figure 2. Probability of $Resp_{\text{Correct}}$ grouped by Internet experiment (web) and laboratory experiment (lab). Solid lines represents the results of the Internet experiment and dashed lines represents the results of the laboratory experiment.

Environment was added which can have the values 'web' or 'lab' (described in Table 5). The second logit regression model reveals that there are no significant differences ($p = 0.174$) between the two experiments. The low average marginal effect of -0.0184 also confirms that the type of the conducted experiments is applicable for an Internet environment.

Figure 2 depicts the mean probability for correctly estimating the number of instruments grouped by the environment. Since in [7] the differences between musicians and non-musicians were emphasized, their data is also depicted in Figure 2. As the logit regression model indicated, the difference in the performance of the participants between the Internet experiment and the laboratory experiment are very low. The participants in the laboratory experiment were about 4.6% better in average for all stimuli than the participants of the Internet experiment. When looking into

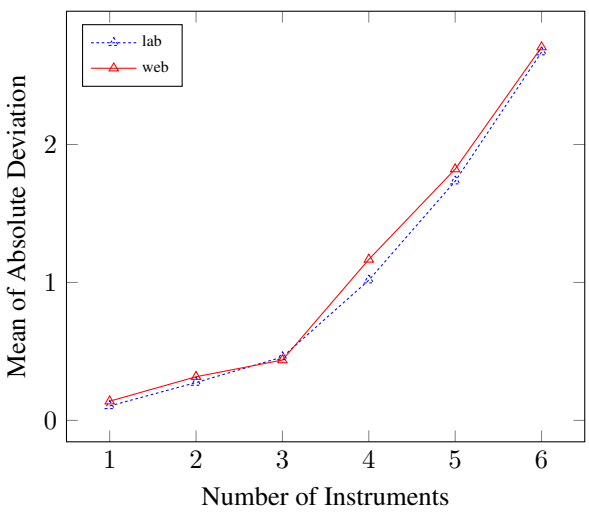

Figure 3. The mean of the absolute deviation grouped by Internet Experiment (web) and laboratory experiment (lab).

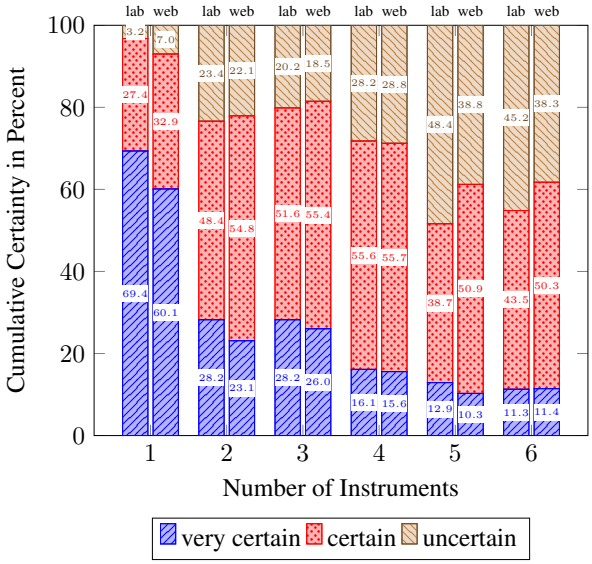

Figure 4. Differences in certainty between the Internet experiment (web) and laboratory experiment (lab).

the differences between musicians and non-musicians, the outcome for the Internet experiment and laboratory experiment differ slightly. In the laboratory experiment musicians performed about 31.6% better than non-musicians and in the Internet experiment musicians performed about 20.85% better.

The probability of correctly estimating the number of instruments does not consider how close an estimation is to the actual number of instruments. This means that a participant who estimates wrong by one instrument for all items has the same $Resp_{\text{Correct}}$ like a participant who is always wrong by two instruments. Despite this work focuses on the correct estimation, the differences in the absolute deviation to the correct number of instruments was also analyzed. The absolute deviation is defined as

$$Dev_{\text{Abs}} = |Num_{\text{Inst}} - Resp|. \quad (2)$$

Figure 3 depicts the mean of Dev_{Abs} for the laboratory experiment and the Internet experiment. A knee point can be seen for $Num_{\text{Inst}} = 3$ where the slope of Dev_{Abs} changes.

To confirm the marginal differences between the Internet experiment and laboratory experiment for Dev_{Abs}, a linear regression model was calculated (see Table 6). Compared to the predictor variable Num_{Inst}, the coefficient of *Environment* is very low.

Another response variable that was obtained from the participants was the certainty of their estimation. Figure 4 depicts the certainty values for the Internet experiment and the laboratory experiment.

For testing whether the environment has a significant influence on the certainty of the participants (*Certainty*), a cumulative link model (also called ordered regression model) was calculated [4]. The cumulative link model is used since *Certainty* is an ordered dependent variable with the possible values 'uncertain', 'certain' and 'very certain'. The predictor variables for the ordered regression model

Coefficient	Estimate	Std. Error	z-value	p-value
(Intercept)	0.08535	0.02706	3.154	0.00161
$Num_{\text{Inst}} = 2$	0.17683	0.01881	9.401	$<$ 2e-16
$Num_{\text{Inst}} = 3$	0.30122	0.01881	16.014	$<$ 2e-16
$Num_{\text{Inst}} = 4$	1.02154	0.01881	54.309	$<$ 2e-16
$Num_{\text{Inst}} = 5$	1.67967	0.01881	89.297	$<$ 2e-16
$Num_{\text{Inst}} = 6$	2.56667	0.01881	136.453	$<$ 2e-16
Environment = '*web*'	0.05481	0.02482	2.208	0.02724

Residual standard error: 0.6597 on 14753 degrees of freedom
Multiple R-squared: 0.6603, Adjusted R-squared: 0.6602
F-statistic: 4779 on 6 and 14753 DF, p-value: $<$ 2.2e-16

Table 6. Linear regression model for Dev_{Abs} calculated from the data obtained by the Internet experiment and the laboratory experiment.

are Num_{Inst} and *Environment*. In Table 7 is the cumulative link model for *Certainty* described. Same as $Resp_{\text{Correct}}$, the environment of the experiment has no significant influence on the dependent variable *Certainty*. Considering the number of participants and the comparable low coefficient, the environment had a very low influence on *Certainty*.

5. DISCUSSION

Regarding the ability of estimating the number of instruments, the web experiment confirmed the results of the laboratory experiment [7]. Both experiments share the same outcome: The probability to correctly estimate up to about three instruments is higher than 50%.

In our previous result analysis of the laboratory experiment [7] we set the focus on the differences between musicians and non-musicians. Between the Internet experiment and the laboratory experiment, slightly different results were obtained when looking into how musicians and non-musicians performed (Figure 2). In the laboratory experiment musicians performed much better compared to non-musicians than in the Internet Experiment. One reason seems to be that in the laboratory experiment 74.2%

Coefficient	Estimate	Std. Error	z-value	p-value
$Num_{\mathrm{Inst}} = 2$	-1.61273	0.05744	-28.078	$<$ 2e-16
$Num_{\mathrm{Inst}} = 3$	-1.43164	0.05701	-25.114	$<$ 2e-16
$Num_{\mathrm{Inst}} = 4$	-2.02315	0.05804	-34.858	$<$ 2e-16
$Num_{\mathrm{Inst}} = 5$	-2.47933	0.05901	-42.013	$<$ 2e-16
$Num_{\mathrm{Inst}} = 6$	-2.43573	0.05900	-41.283	$<$ 2e-16
Environment = 'web'	-0.01008	0.07355	-0.137	0.891
Threshold coefficients:				
	Estimate	Std. Error	z-value	
$uncertain\|certain$	-2.90465	0.08434	-34.441	
$certain\|verycertain$	-0.41072	0.08090	-5.077	
AIC: 28242.52				

Table 7. Logit cumulative link model of certainty that was calculated from the data obtained by the Internet experiment and the laboratory experiment.

of the musicians had also a professional background in audio. In the Internet experiment only 27.5% of the musicians had a professional background in audio. Since audio professionals more often have to detect hardly audible differences in audio files they are more trained in this field. As mentioned before, being a musician in our context does just mean that the participant plays an instrument without any information about his expert level or the time he or she spends on practicing.

When examining the responses for items with the same number of instruments being played back, noticeable differences between the genres were found. For the stimuli with $Num_{\mathrm{Inst}} \leq 4$ the mean of $Resp_{\mathrm{Correct}}$ was 53.0% and for non-classical items 62.5%. Since the experiment was not designed for testing classical versus non-classical items, we cannot make definitive statements about whether humans are better in estimating instruments for a specific music genre. Moreover, we did not address this issue in our result analysis.

One of the main reasons for conducting the experiment was to find out which influence the Internet environment has on the results. All three regression models which included data of both experiments (Table 5, 6 and 7) revealed that the Internet environment had only very minor effects on the results. Moreover, despite the large number of participants in both experiments, the predictor variable *Environment* was even not significant in two out of three regression models.

It is often recommended to use headphones instead of loudspeakers for Internet experiments. From the relative low average marginal effect of *Setup* (Table 4), it can be derived that the type of the setup had only minor effects on the results of the Internet experiment.

Surprising was the small number of participants who had to be screened. We excluded 27 participants since they gave at least one non-serious response which is about 2.3% (1195 participants remained after excluding all trials which were not the first ones). Most of these excluded participants responded with either very high numbers (e. g. 99) or responded with zeros for their estimated number of instruments being played. We assume that especially participants who responded with zero only wanted to get an impression of the experiment.

6. CONCLUSION

The Internet experiment presented in this paper confirmed the results of a previous laboratory experiment that humans are able to correctly estimate up to around three instruments in music. Furthermore significant differences in performance between musicians and non-musicians found out by the previous experiment were confirmed. The comparison between the results of the Internet experiment and laboratory experiment revealed that only minor differences between both environments exist. Using headphones instead of loudspeaker is often held to be important when conducting listening tests over the Internet. In this experiment the audio setup used had only a minor influence on the results. According to these results, experiments in the fields of music perception or music information retrieval related to our procedure are well suited for being conducted over the Internet.

7. REFERENCES

[1] Albert S. Bregman. *Auditory Scene Analysis: The Perceptual Organization of Sound*. Bradford Books, MIT Press, Cambridge, 1990.

[2] Albert. S. Bregman and Jeffrey Campbell. Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of experimental psychology*, 89(2):244–9, August 1971.

[3] Emilios Cambouropoulos. Voice and stream: Perceptual and computational modeling of voice separation. *Music Perception*, 26(1):75–94, 2008.

[4] Rune Haubo B. Christensen. Analysis of ordinal data with cumulative link models – estimation with the R-package ordinal, 2012.

[5] EBU. Loudness normalisation and permitted maximum level of audio signals (EBU Recommendation R 128), 2011.

[6] Peter Eckersley. How Unique Is Your Web Browser? In *Privacy Enhancing Technologies Symposium (PETS 2010)*, pages 1–18, Berlin, Germany, 2010.

[7] Fabian-Robert Stöter and Michael Schoeffler and Bernd Edler and Jürgen Herre. Human ability of counting the number of instruments in polyphonic music. volume 19, page 035034. ASA, 2013.

[8] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: Popular, classical, and jazz music databases. In *Proc. ISMIR*, pages 287–288, 2002.

[9] David Huron. Voice Denumerability in Polyphonic Music of Homogeneous Timbres. *Music Perception: An Interdisciplinary Journal*, 6(4):361–382, 1989.

[10] Jin Ha Lee. Crowdsourcing music similarity judgments using mechanical turk. In *ISMIR*, pages 183–188, 2010.

[11] Ulf-Dietrich Reips. Standards for Internet-Based Experimenting. *Experimental Psychology*, 49(4):243–256, 2002.

[12] Ulf-Dietrich Reips. Using the Internet to Collect Data. In Harris Cooper, Paul M. Camic, Debra L. Long, A. T. Panter, David Rindskopf, and Kenneth J. Sher, editors, *APA Handbook of Research Methods in Psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological*, volume 2, chapter 17, pages 291–310. American Psychological Association (APA), Washington, US, 2012.

[13] Matthew J. Salganik, Peter Sheridan Dodds, and Duncan J. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *Science (New York, N.Y.)*, 311(5762):854–6, February 2006.

[14] Catherine J. Stevens. Music perception and cognition: A review of recent cross-cultural research. *Topics in cognitive science*, 4(4):653–667, 2012.

[15] D. Wang and G. Brown. *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE Press, 2006.

[16] Norma Welch and John H. Krantz. The World-Wide Web as a medium for psychoacoustical demonstrations and experiments: Experience and results. *Behavior Research Methods, Instruments, & Computers*, 28(2):192–196, June 1996.