Perceptual evaluation of violins: A quantitative analysis of preference judgments by experienced players

Charalampos Saitis[a)]
Computational Acoustic Modeling Laboratory, Centre for Interdisciplinary Research in Music Media and Technology, Schulich School of Music, McGill University, Montreal, Quebec H3A 1E3, Canada

Bruno L. Giordano
Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, Scotland, United Kingdom

Claudia Fritz
Lutheries-Acoustique-Musique, Institute Jean le Rond d'Alembert, Université Pierre et Marie Curie, UMR CNRS 7190, 4 place Jussieu, 75005 Paris, France

Gary P. Scavone
Computational Acoustic Modeling Laboratory, Centre for Interdisciplinary Research in Music Media and Technology, Schulich School of Music, McGill University, Montreal, Quebec H3A 1E3, Canada

The overall goal of the research presented here is to better understand how players evaluate violins within the wider context of finding relationships between measurable vibrational properties of instruments and their perceived qualities. In this study, the reliability of skilled musicians to evaluate the qualities of a violin was examined. In a first experiment, violinists were allowed to freely play a set of different violins and were then asked to rank the instruments by preference. Results showed that players were self-consistent, but a large amount of inter-individual variability was present. A second experiment was then conducted to investigate the origin of inter-individual differences in the preference for violins and to measure the extent to which different attributes of the instrument influence preference. Again, results showed large inter-individual variations in the preference for violins, as well as in assessing various characteristics of the instruments. Despite the significant lack of agreement in preference and the variability in how different criteria are evaluated between individuals, violin players tend to agree on the relevance of sound "richness" and, to a lesser extent, "dynamic range" for determining preference. © 2012 Acoustical Society of America. [http://dx.doi.org/10.1121/1.4765081]

I. INTRODUCTION

The quality of a violin depends on a number of factors, many of which relate directly to the sound radiated by the instrument, as well as others that relate to the interaction between the player and the instrument. For example, an important aspect of a violin's behavior concerns its playability or response to various playing gestures (Woodhouse, 1993a). Some of this information may be communicated to the player via tactile and proprioceptive channels (e.g., hands, arms, chin). In the current study, we adopted a playing-based evaluation approach to investigate the perceptual processes involved when the player compares different violins in a musical setting—e.g., during the process of choosing a new instrument. As a starting point, we focused on the key question of how consistent experienced players are at assessing violins and whether there is agreement between violinists.

Attempts to quantify the characteristics of "good" and "bad" violins from listening tests and/or acoustical and structural dynamics measurements have largely been inconclusive.

Willgoss and Walker (2007) carried out semantic differential tests on recorded samples from 12 Stradivari and Guarneri violins performed by the same player. The recordings were judged by two independent groups of listeners, 15 university-level music students and 8 professional musicians, based on 13 bipolar pairs of verbal timbre descriptors (e.g., *resonant-muffled*). Both participant groups showed little or no agreement, with professional violinists appearing more self-consistent than students (no quantitative data are provided).

Fritz *et al.* (2007) carried out a series of listening tests, whereby recorded bridge force signals were convolved with measured and post-processed bridge admittances to synthesize violin sounds, allowing controlled variations of modal properties. Initial tests demonstrated a large variability in thresholds for the perception of acoustical modifications across participants, with significantly lower thresholds for experienced musicians than for subjects with little musical training. Subsequent tests showed that when listening to recorded single notes with varying levels of vibrato and body damping, ratings on the perceived *liveliness* of the sound were inconsistent across participants, while overall preference judgments appeared to be consistent (Fritz *et al.*, 2010). Interestingly, when asked to play freely on an electric

[a)]Author to whom correspondence should be addressed. Electronic mail: charalampos.saitis@mcgill.ca

violin (i.e., the bridge force signal was passed through the modified admittances in real time), participants rated both *liveliness* and preference consistently. In another set of tests, participants were found consistent at assessing virtual violin sounds described as *bright, harsh*, and *clear,* but less so for *nasal* and *good* (Fritz *et al.*, 2012). Subsequent analyses suggested that the observed inter-individual variability might have resulted from the fact that different players evaluate different qualities of the violin in different ways. We further investigated this hypothesis in the current study.

Listening tests using recordings, synthesized sounds or live performance have several disadvantages. Recorded sounds often lack the naturalness of live performance. Similarly, synthesized tones often sound rather unmusical (Wright, 1996). And when using live players, listeners, regardless of musical relevance or the lack thereof, tend to focus more on the skills of the performer than the instrument. Most importantly, it is virtually impossible to assess vibro-mechanical properties without direct interaction with the instrument. Concerning the perspective of the player, listening tests are therefore not completely indicative of the processes that take place when assessing the qualities of a violin; playing-based evaluations afford a higher level of ecological validity. By playing, violinists can experience a wider range of performance effects than the very short phrases or single notes often used in listening tests, and in this way assess any particular attribute of the instrument based on multi-modal sensory data (i.e., based on auditory and tactile feedback).

Several studies have attempted to correlate mechanical characteristics to instrument quality. Alonso Moral and Jansson (1982) suggested the importance of the signature modes below 600 Hz and the bridge hill in the 2–3 kHz range on violin sound quality based on bridge admittance measurements on 24 violins, which had previously been played and rated on tonal quality by two professional violinists. Bissinger's wide range of vibration and radiation measurements on 17 violins with quality ratings from bad to excellent, provided by a professional player (12 violins) and Bissinger himself (5 violins), showed no significant quality differentiators except for the Helmholtz-like cavity mode A0, the radiation of which was significantly stronger for excellent than for bad violins (Bissinger, 2008). It is unclear whether the results of these studies are reliable or generalizable, primarily because the evaluation tasks were carried out with an extremely low number of participants. It is also unclear whether the influences of parameters like the choice of bow or visual information (e.g., varnish, identity of the instrument) were controlled because these specifics were not published. Attempts to correlate measurable vibrational properties of the violin with perceptual judgments by players first require a closer look into the subjective evaluation process.

In this study, we investigated the perceptual evaluation of violins from the player's perspective by focusing on preference judgments by experienced violinists. To this end we designed two violin-playing perceptual tests. In experiment 1, participants were asked to rank a set of different violins from least to most preferred, and provide rationale for their choices. We investigated intra-individual consistency and inter-individual

agreement in the preference rankings. From the verbal data, we extracted attributes of the violin that were then used to structure rating scales for a subsequent study. In experiment 2, participants were asked to rate a different set of violins according to specific attributes and preference. We investigated the origin of inter-individual differences in the preference for violins and measured the extent to which different attributes of the instrument are associated with preference.

II. EXPERIMENT 1

The goal of this experiment was to examine how consistent skilled players are at assessing violins within and between themselves. We investigated intra-individual consistency and inter-individual agreement over repetitive preference rankings by skilled violin players. Preference judgments were collected as a measure of subjective evaluation based on choice behavior (Giordano *et al.*, 2012). We asked participants to provide rationale for their choices through a specially designed questionnaire and extracted attributes of the violin that were then used to design the rating scales for experiment 2.

A. Method

1. Participants

Twenty skilled string players took part in this experiment (8 females, 12 males; average age = 34 years, SD = 13 years, range = 20–65 years; 11 native English speakers, 3 native French speakers, 6 other). They had at least 15 years of violin experience (average years of violin training = 26 years, SD = 12 years, range = 15–60 years; average hours of violin practice per week = 15 h, SD = 9 h, range = 9–30 h), owned violins with estimated prices ranging from less than $1000 to $30 000, and were paid for their participation. Thirteen participants described themselves as professional musicians, and 8 had higher-level degrees in music performance (MMus, MA, DMus, DMA). They reported playing a wide range of musical styles [classical (95%), folk (47%), baroque (37%), jazz/pop (10%), and contemporary (5%)] and in various types of ensembles [chamber music (70%), symphonic orchestra (70%), solo (55%), and folk/jazz band (40%)].

2. Violins

Eight violins of different make (Europe, North America, China), age (1840–2010), and price ($1000-$65 000) were used (see Table I). They were chosen from several local luthier workshops in order to form, as much as possible, a set of violins with a wide range of characteristics. The violins had not been played on a regular basis, two having been recently fabricated and most of the others coming from the available sales stock of a workshop. Student-level violins were not used because skilled players consistently discriminated these from performance-level instruments in a pilot study.[1] Participants' own violins were not included in the set of instruments in order to avoid possible preference biases caused by the mere exposure effect (Zajonc, 1968) by which familiarity with a stimulus object increases preference toward it. The respective luthiers provided the price estimates and tuned the instruments for optimal playing condition based on their

TABLE I. Violins used in experiment 1 in order of price along with preference score averaged across participants.[a]

Violin	A	B	C	D	E	F	G	H
Origin	France	Italy	Canada	Canada	Canada	Germany[b]	France	China
Luthier[c]	Silvestre	Cavallini	…	…	…	Unknown	Apparut	…
Year	1840	1890	2010	2010	1976	Unknown	1936	2010
Price	$65K	$35K	$16K	$13K	$10K	$8K	$6K	$1.3K
Preference[d]	0.51(0.05)	0.44(0.05)	0.44(0.06)	0.37(0.07)	0.44(0.06)	**0.54(0.05)**	0.42(0.02)	*0.34(0.04)*

[a]0 = never preferred to any other violin; 1 = always preferred to all other violins; standard error of the mean in parentheses.
[b]This is based on a luthier's informal appraisal, as there is no information regarding the make and age of this violin.
[c]The names of living luthiers are not provided for confidentiality purposes.
[d]The most preferred violin (F) is indicated in bold and the least preferred violin (H) in italics.

own criteria. Participants were given the option to either use a provided shoulder rest (Kun Original model), or use their own, or use no shoulder rest. The fact that some violins may have been less optimally tuned or had strings of varying quality was not a concern, as that should not have influenced the consistency of the preference rankings.

3. Controls

Anecdotal evidence strongly suggests that some visual information, such as the color of the varnish, the grain of the wood, or identifying marks of the violin, may influence judgment. More specifically, possible recognition of the instrument's make and origin is likely to produce preference biases (e.g., old Cremonese violins are often considered excellent and hence preferred over modern instruments). To help minimize the effects of such visual cues as much as possible in listening tests involving live performance, the listeners or the performers or both are often blindfolded. Another approach is to have the instruments played behind a physical divider (e.g., Petiot and Causse, 2007). However, blindfolding was not a viable solution for our playing tests because players were allowed to freely explore a set of violins and rank-order them on a table. To circumvent the potential impact of visual information on judgment while ensuring a certain level of comfort for the musicians, as well as safety for the instruments, we used low light conditions and asked participants to wear dark sunglasses. Based on this procedure, violinists could provide unbiased assessments while still retaining some visual contact with the instruments.

A critical issue when conducting violin playing tests is the choice of a bow. In the present study, two options were considered: using a common bow across all participants (e.g., Inta et al., 2005) or asking players to use their own bow. Although neither solution is ideal, by considering the bow as an extension of the player (second option), we avoided the potential problems of using a common bow (e.g., participants being uncomfortable with a bow they are not familiar with). Further, a common bow would potentially trigger a similar quality debate (Caussé et al., 2001). Having the participants use the bow that they are most familiar with was also felt to be more representative of how violinists assess instruments while in the process of purchasing one.

The experimental sessions took place in a diffuse room with a surface of 27 m^2 and reverberation time of ~0.18 s to help minimize the effects of room reflections on the direct sound from the violins (Bissinger and Gearhart, 1998).

4. Procedure

The experimental session lasted 2 h and was organized in two phases. The experimenter was constantly present in the room to facilitate the procedure. In the first phase, participants were presented with the violins randomly ordered on a table by the experimenter. They were asked to play all instruments for up to 25 min in order to familiarize themselves with the set. The second phase consisted of five trials. On each trial of the second phase, participants were initially presented with all violins placed on a table in random order (determined by computer calculations) by the experimenter. They were then given up to 15 min to play, evaluate, and rank the violins by placing them in order of preference (from least to most preferred) on a different table. Participants were not allowed to assign the same preference rank to two or more instruments. Rankings were recorded by the experimenter. Participants were instructed to maximize evaluation speed and accuracy. No playing constraint was imposed on the evaluation process (e.g., specific repertoire). Participants were instead instructed to follow their own strategy. They were encouraged to play their own violin whenever they needed a reference point during the experiment. To minimize fatigue, participants were encouraged to take breaks between trials whenever needed. Upon completing the first trial, participants provided free verbal (written) responses to the question "How and based on which criteria did you make your ranking?" At the end of each subsequent trial, they were given the opportunity to modify their initial response if they so wanted. Upon completing the last trial, participants provided written responses to the questions "Do you have any comments or remarks about the task you were involved in?" and "To what extent was wearing sunglasses disturbing?" Participants were asked to return for a second, identical session 3–7 days after having completed the first session. In total, participants ranked each violin $5 \times 2 = 10$ times.

B. Results

We carried out four different analyses. First, we measured and analyzed the levels of intra- and inter-individual consistency in the preference rankings. Second, we assessed

the extent to which various characteristics of the participants explained their ability to be consistent across repeated preference-ranking trials (e.g., whether "hours of practice per week" was correlated with self-consistency). Third, we derived an overall measure of preference for each of the violins, and assessed differences in preference across violins. Finally, we analyzed the verbal descriptions of the violin attributes relevant in determining the preference responses given by the participants.

1. Intra- and inter-individual consistency

Consistency was measured as the concordance correlation between preference rankings from different trials. The concordance correlation ρ_c is a special case of the Pearson product-moment correlation coefficient that measures departures from the equality lines with slopes $\pm 45°$: $\rho_c(A, B) = 1$ and -1 if $A = B$ and $A = -B$, respectively, and $\rho_c(A, B) = 0$ in case of no association between A and B (Lin, 1989). The concordance correlation coefficient is appropriate for measuring the agreement between both continuous (e.g., ratings) and ordinal level data (e.g., preference rankings) (Shoukri, 2004). The first step involved computing a 200×200 symmetric matrix of ρ_c coefficients between the rankings on each of the 10 trials for each of the 20 participants. Across the 19 900 cells of the lower triangular part of this correlation matrix, there were 19 000 correlations between trials from different participants and 900 correlations between trials from the same participant. Across the 900 correlations between rankings from the same participant, 500 correlations are between trials from different sessions and 400 correlations are between trials from the same session. Figure 1 displays the histograms for all the ρ_c measures computed between preference rankings from the same participant, and between preference rankings from different participants,
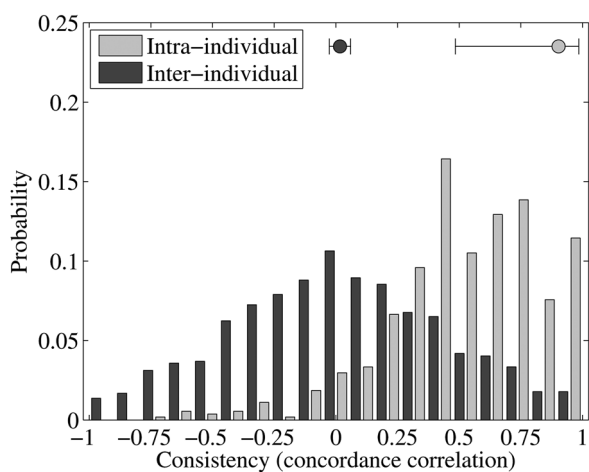

FIG. 1. Experiment 1: Distribution of intra- and inter-individual concordance correlation coefficients, computed between violin-preference rankings from the same and different participants, respectively: 1 corresponds to perfect consistency, 0 corresponds to no consistency, -1 corresponds to perfect anti-consistency (i.e., exactly opposite rankings given on different trials). The symbols above the histograms report the across-participants average of the intra- and inter-individual consistency scores (0.903 and 0.017, respectively; error bar = 95% confidence interval of the mean; the ordinate for the symbols has been chosen arbitrarily for display purposes). See the text for details on averaging of concordance correlations.

respectively. The intra-individual ρ_c distribution is highly asymmetrical with peaks in the range 0.5–0.8, whereas the inter-individual ρ_c distribution is roughly symmetrically centered around zero. In order to give a preliminary approximate figure for the results of this analysis, an initial test assessed how many of these ρ_c coefficients were significant when assuming their independence ($p < 0.05$, $df = 6$): the percentage of significant ρ_c coefficients between rankings from the same participants and between rankings from different participants was 51% and 7%, respectively. The first of these figures corresponds, approximately, to the case where the consistency between all of the 10 rankings given by the same participant throughout the experiment was significant for 10 of the 20 participants (51% of the intra-individual ρ_c coefficients). The second of these figures corresponds, approximately, to the case where all of the rankings from different participants in a group of 6 out of the 20 participants were significantly consistent with each other (the number of ρ_c coefficients between the trials of two different participants equals 100; the number of ρ_c coefficients between the trials of 6 different participants equals 1500, i.e., 7.89% of all the inter-individual ρ_c coefficients between the trials of all of the 20 participants).

Further, more rigorous analyses were carried out on measures of intra- and inter-individual consistency computed for each of the participants. The intra-individual consistency was given by the average of the ρ_c between the preference rankings from each of the 10 trials for the same participant. The computation of the inter-individual consistency for a given participant A was given by the average of the ρ_c measures between the rankings of A and the rankings of all of the other participants. Note that according to this definition, the inter-individual consistency measures for participants A and B would be computed by considering the same set of 100 ρ_c measures between the 10 rankings of participant A and those of participant B. In order to minimize issues of nonindependence between the inter-individual consistency measures for different participants, correlations were equally distributed among participants at random (e.g., for participant A the inter-individual consistency measure considered 50 randomly selected $\rho_c(A, B)$ measures, whereas for participant B it included the other 50). On average, whereas the measures of intra-individual consistency were significantly higher than zero, average value = 0.903 [$t(19) = 3.24$, $p = 0.004$], the measures of inter-individual consistency were not significantly different than zero, average value = 0.017, [$t(19) = 0.79$, $p = 0.439$]. Similarly, the measures of intra-individual consistency were significantly higher than those of inter-individual consistency [paired sample $t(19) = 3.24$, $p = 0.004$]. Figure 1 reports the intra- and inter-individual consistency measures averaged across participants (see symbols above the histograms).[2]

The same methodology was adopted to carry out a more detailed analysis of the variation of intra- and inter-individual consistency across the two experimental sessions. For both experimental sessions, the average measure of intra-individual consistency was significantly higher than zero, average value = 0.947 and 0.963 for sessions 1 and 2, respectively [$t(19) \geq 2.88$, $p \leq 0.035$], and the average measure of inter-individual consistency was not significantly

different than zero, average value $= -0.004$ and 0.045 for sessions 1 and 2, respectively [$t(19)) \leq 1.84$, $p \geq 0.081$]. For both experimental sessions, the average intra-individual consistency was significantly higher than the average inter-individual consistency [paired sample $t(19) \geq 2.23$, $p \leq 0.037$]. Finally, whereas intra-individual consistency did not significantly differ between sessions 1 and 2 [paired sample $t(19) = -0.26$, $p = 0.800$], the inter-individual consistency was significantly higher in session 2 than in session 1 [paired sample $t(19) = -2.67$, $p = 0.015$]. Note, however, that the increase in inter-individual agreement from session 1 to session 2 is negligible because it corresponds to an increase in the average of the inter-individual ρ_c measure of 0.050.

2. Influence of participant characteristics

We assessed whether known characteristics of the participants explained the variability across participants in intra-individual consistency. A two-sample t-test was adopted to assess whether intra-individual consistency significantly differed between professional and amateur violin players ($N = 13$ and 7, respectively). Despite a tendency for professional violin players to be more self-consistent than amateur players, average intra-individual consistency $= 0.948$ and 0.704, respectively, the difference proved to be not significant [independent samples $t(18) = 0.98$, $p = 0.209$, unequal variance]. We then computed the Spearman rank correlation ρ_S between measures of intra-individual consistency on the one hand, and the self-reported price of the owned violin, the years of violin training, and the weekly hours of violin practice, on the other. Average imputation was used to replace missing values for these self-reported measures. None of the correlations was significant, $\rho_S \leq 0.272$ ($p \geq 0.245$, $df = 18$].

3. Preference ranking of the violins

For each participant, and for each of the violins, we computed a preference score defined as the proportion of times that a violin was ranked as preferred to all of the other violins throughout all the preference-ranking trials. The across-participants average preference scores for each violin are reported in Table I and plotted in Fig. 2.

4. Verbal descriptions of violin attributes

Finally, we examined the spontaneous verbal responses of participants to the question "How and based on which criteria did you make your ranking?" A total of 194 phrasings were coded into violin attributes and classified according to whether they described the sound (e.g., *richness*), the instrument (e.g., *weight*), or the interaction between the player and the instrument (e.g., *easy to play*). Class-attribute pairs (e.g., sound-*richness*) that were reported multiple times by the same participant across different trials and/or sessions were considered only once. A total of 95 attributes of the violin were thus extracted (see Table II). The sole purpose of this analysis was to extract those attributes of the violin that participants considered important for preference in order to facilitate the design of attribute-rating scales for experiment
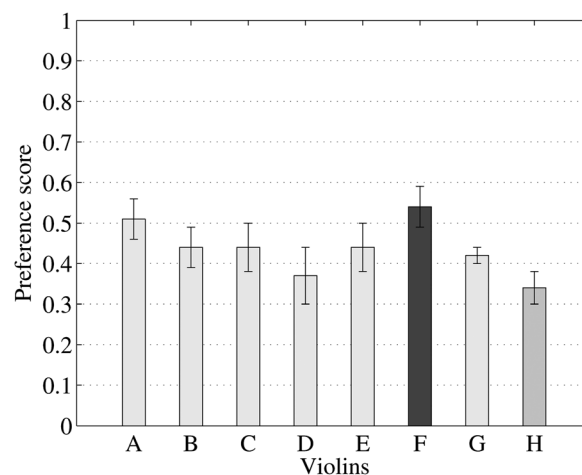

FIG. 2. Experiment 1: Across-participants average of the preference score for each violin (error bar = 95% confidence interval of the mean). The violins are ordered by decreasing price. The darkest bar indicates the most preferred violin (F); the less dark bar indicates the least preferred violin (H). See the text for details on computing of scores.

2 (see Sec. III A 3). More comprehensive analyses of the verbal data collected in this study will be discussed in a separate paper.

TABLE II. Number (N) of occurrences across participants of free verbal descriptions[a] for violin preference ranking criteria based on the analysis of the verbal data collected in experiment 1.[b]

	Interaction	Sound	Instrument	N
Easy to play	X			**11**
Response	X			**8**
Resonance		X		7
Richness		X		**7**
Projection		X		6
Weight			X	4
Playability	X			4
Clarity		X		4
Color		X		4
Feel	X			3
Beauty		X		3
Consistency		X		*3*
Evenness		X		*3*
Power		X		3
Problems		X		3
Shape			X	2
Balance (sound)		X		2
Balance (response)			X	2
Equality		X		2
Comfort	X			2
Flexibility	X			2
Depth		X		2
High register		X		2
Overtones		X		2
Speaks		X		2
Strong		X		2
Sustained		X		2

[a]Descriptions with fewer than two occurrences are not included.
[b]For the purposes of experiment 2, only those attributes mentioned by at least five participants were considered (above the horizontal line) and only the three indicated in bold used. The various verbalizations semantically related to "balance" (across the strings) are indicated in italics.

C. Discussion

The results of this experiment showed that experienced violinists are self-consistent when assessing different instruments in terms of preference both within and across different-day experimental sessions. Despite 15 participants reporting in the questionnaire that the task was difficult, intra-individual consistency was high overall. Further, only four players reported being bothered by the dark sunglasses. However, the various analyses reported previously demonstrated a significant lack of agreement between string players in the preference for violins. Further, attempts to associate self-consistency with known (self-recorded) characteristics of the participants were rather inconclusive. In particular, there were no significant differences in self-consistency between professional violin players and amateur musicians, which appear to contrast with previous observations in listening tests (Willgoss and Walker, 2007). Finally, we observed no effect of training from session 1 to session 2 on self-consistency. Interestingly, violinists were not significantly more self-consistent within one experimental session than across multiple sessions carried out in different days. This result suggests that the criteria used by individuals to evaluate violin preference remain relatively stable within a short time span.

The large inter-individual differences observed in the preference for violins could have two different origins. First, individual violin players may disagree on what particular qualities they look for in a violin. For example, some violinists may have a strong preference for violins that produce bright tones irrespective of differences in other sound or vibrational characteristics, whereas others may favor instruments that are easy to play notwithstanding how bright the resulting tone is. Similarly, the fact that the participants were using their own violins as a reference during the rankings could have exaggerated this effect. Second, different violin players may follow different processes to assess those qualities considered essential for the evaluation of an instrument (Fritz *et al.*, 2010). For example, all violinists may prefer instruments that are easy to play but there may be differences in how ease of playing is evaluated across individuals. To tease apart these potential sources of variation across players in the preference for violins, we carried out a subsequent experiment to examine whether there would be more inter-individual agreement if violinists are asked to focus on specific attributes of the instrument.

III. EXPERIMENT 2

The goals of this experiment were to investigate the origin of the large inter-individual differences in the preference for violins observed in experiment 1 and measure the extent to which different attributes are associated with preference. As well, we were interested to know how consistency would be affected if subjects were asked to focus on particular violin attributes when considering preference. We investigated intra-individual consistency and inter-individual agreement over repetitive ratings by experienced players on specific attributes of the violin as well as preference. The rating scales were determined based on the analysis of verbal data

collected in experiment 1 as well as the potential for the descriptors to be correlated with measured vibrational properties of the violin.

A. Method

1. Participants

Thirteen skilled string players took part in this experiment (9 females, 4 males; average age = 28 years, SD = 9 years, range = 21–53 years; 8 native English speakers, 4 native French speakers, 1 other). They had at least 15 years of violin experience (average years of violin training = 22 years, SD = 9 years, range = 15–46 years; average hours of violin practice per week = 25 h, SD = 11 h, range = 7–42 h), owned violins with estimated prices ranging from \$3000 to \$30 000, and were paid for their participation. Eleven participants described themselves as professional musicians, and six had higher-level degrees in music performance (MMus, MA, DMus, DMA). They reported playing a wide range of musical styles [classical (92%), folk (31%), baroque (23%), and jazz/pop (15%)] and in various types of ensembles [chamber music (92%), symphonic orchestra (85%), solo (85%), and folk/jazz band (31%)].

2. Violins

Ten violins of different make (Europe, North America, China), age (1770–2009), and price (\$2000-\$250 000) were used (see Table III). They were chosen from several local luthier workshops in order to form, as much as possible, a set of violins with a wide range of characteristics. The violins had not been played on a regular basis as most were from the available sales stock of the workshops. One of the violins (H) had been investigated in experiment 1 (the most preferred, labeled F in Table I). Similarly to the previous experiment, student-level violins and the participant's own violin were not included in the set of instruments, and the respective luthiers provided the price estimates and tuned the instruments for optimal playing condition based on their own criteria. Participants were given the option to either use a provided shoulder rest (Kun Original model), or use their own, or use no shoulder rest. All other experimental conditions (i.e., visual occlusion, choice of a bow, and room) were as in experiment 1.

3. Criteria

In view of logistical constraints (i.e., duration of the experimental session), we chose to consider only those attributes of the violin mentioned by at least 25% of the participants in experiment 1 (see Table II). From these, *resonance* and *projection* were discarded due to potential problems associated with their evaluation in the present experimental context. For example, sound projection is a difficult quality to judge reliably solely by playing the violin (Loos, 1995). We then decided to add a *balance* (response across the strings) rating scale because we noticed that several violinists used verbalizations that were semantically related to this attribute (e.g., *evenness, consistency, equality*) (see Table II). Even though not justified by the analysis of the verbal data, we

TABLE III. Violins used in experiment 2; also reported are the across-participants average of the preference, sound richness, and dynamic range scores.[a]

Violin	A	B	C	D	E	F	G	H	I	J
Origin	Italy	Italy	Germany	Italy	France	France	France	Germany[b]	Canada	China
Luthier[c]	Gagliano	Storioni	Fisher	Sderci	Kaul	...	Guarini	Unknown	...	...
Year	1770–1975	1799	1787	1964	1933	2009	1877	Unknown	2005	2006
Price ($1000)	250	44	22	20	20	17	11	8	6	2
Preference[d]	0.41(0.05)	0.47(0.07)	**0.62(0.07)**	0.33(0.07)	0.53(0.07)	0.47(0.05)	0.32(0.05)	0.51(0.06)	*0.28(0.05)*	0.55(0.05)
Richness	0.37(0.06)	0.47(0.07)	0.59(0.06)	0.29(0.06)	0.49(0.08)	0.49(0.05)	0.41(0.04)	0.59(0.06)	0.28(0.04)	0.50(0.06)
Dynamic range	0.45(0.05)	0.41(0.05)	0.53(0.06)	0.42(0.06)	0.61(0.08)	0.39(0.08)	0.41(0.06)	0.40(0.07)	0.37(0.06)	0.50(0.06)

[a]0 = never judged as more preferred, richer, or having a wider dynamic range than any of the other violins; 1 = always judged as more preferred, richer, or having a wider dynamic range than all other violins; standard error of the mean in parentheses.

[b]Based on a luthier's informal appraisal, as there is no information regarding the make and age of this violin.

[c]Names of living luthiers are not provided for confidentiality purposes.

[d]The most preferred violin (C) is indicated in bold and the least preferred violin (I) in italics. Violin H was included in experiment 1 (labeled F, highest preference score.

included *dynamic range* because it has long been a source of investigation in the literature (Askenfelt, 1989; Woodhouse, 1993b; Schoonderwaldt *et al*., 2008). These five criteria had been previously proposed as part of a standardized procedure for evaluating violins (Bissinger and Gearhart, 1998). A very similar set was obtained when Inta *et al*. (2005) asked violinists to report evaluating qualities for purchasing a violin. Finally, we added an overall preference rating scale in order to examine the extent to which each of the selected attributes influences preference.

To ensure common interpretation of the rating scales across all participants as much as possible, each criterion was presented in the form of a descriptive phrase alongside a short explanatory text: (1) *the violin is easy to play* (it requires minimal effort to produce sound, easy to avoid wolf tones, easy to "get around" the instrument); (2) *the violin responds well* (it produces desired sounds using a wide range of bowing gestures, it responds well to a wide range of actions of the player); (3) *the violin has a rich sound* (the violin produces a sound that is rich in harmonics and overtones); (4) *the violin is well balanced across the strings* (the playing behavior of this violin is similar across all strings); (5) *the violin has a broad dynamic range (from piano to forte)* (it can produce sounds of a wide range of dynamics, from piano to forte); and (6) *the violin is the one I prefer the most* (self-explanatory).

For all criteria, unipolar continuous rating scales were preferred over bipolar scales. For the latter it is necessary to use antonyms that are semantically relevant (e.g., *male:female*).

However, considering *poor* as the opposite of *rich* may not be pertinent to evaluating the sound of a violin (Fritz *et al*., 2012). To comply with the descriptive form in which each criterion was presented to participants, the right end of each unipolar scale was labeled as "strongly agree" and the left end was labeled as "strongly disagree" (see Fig. 3).

4. Procedure

The experimental session lasted 2 h and was organized in three phases. In the first phase, participants were presented with the violins and the rating criteria. They were asked to play all instruments for 20 min to acquaint themselves with the set. Participants were also instructed to explore how much each attribute varied across the different violins in the set. The second phase involved a short training session with two trials to help participants familiarize themselves with the rating task. On each trial, participants were presented with a violin, which was not one of the 10 violins used in the main session, and asked to rate it according to the given criteria. In the third phase, each of the 10 violins was presented once in each of three subsequent blocks of 10 trials, for a total of 30 trials. Participants thus rated each violin three times. The order of presentation of the violins within each block of trials was randomized (determined by computer calculations). On each trial, participants were asked to play and rate the violin according to each criterion on a unipolar continuous scale using on-screen sliders. They had to move each slider (i.e., rate each criterion) before being allowed to move
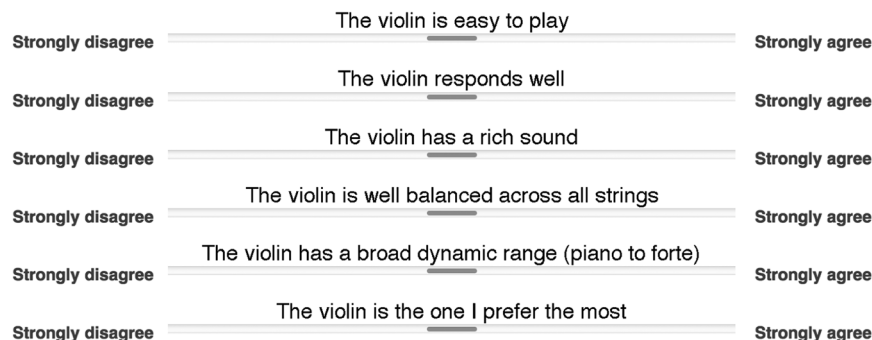

FIG. 3. Testing interface used to collect the ratings in experiment 2.

Saitis *et al*.: Perceptual evaluation of violins

to the next trial (i.e., violin). In order to end a trial and start the succeeding one, the participant clicked an on-screen button labeled "Next" that appeared only after all sliders had been moved. Participants were instructed to maximize evaluation speed and accuracy. No playing constraint was imposed on the evaluation process (e.g., specific repertoire). Participants were instead instructed to follow their own strategy. They were encouraged to play their own violin whenever they needed a reference point during the experiment. To minimize fatigue, they were encouraged to take breaks between each block of trials. Upon completing the last trial, participants provided written responses to the questions "Do you have any comments or remarks about the task you were involved in?" and "To what extent was wearing sunglasses disturbing?"

B. Results

We carried out four different analyses. First, we compared the measures of intra- and inter-individual consistency for each of the rated attributes, and further assessed significant differences between the intra- and inter-individual consistency measures for the preference scale on the one hand, and each of the other attribute-rating scales, on the other. As a part of this analysis, we also compared the measures of intra- and inter-individual consistency recorded during experiment 1 with those recorded during experiment 2 for the preference rating scale. Second, we assessed the effects of participant characteristics on the measures of intra-individual consistency computed for each of the attribute-rating scales. Third, we assessed significant differences between the group-average preference for the different violins. Finally, we measured the extent to which preference ratings for each participant could be predicted based on ratings of the different attributes.

1. Intra- and inter-individual consistency

For each rating scale, intra- and inter-individual consistency was measured and assessed based on the ρ_c between ratings given on different blocks of trials. We followed the
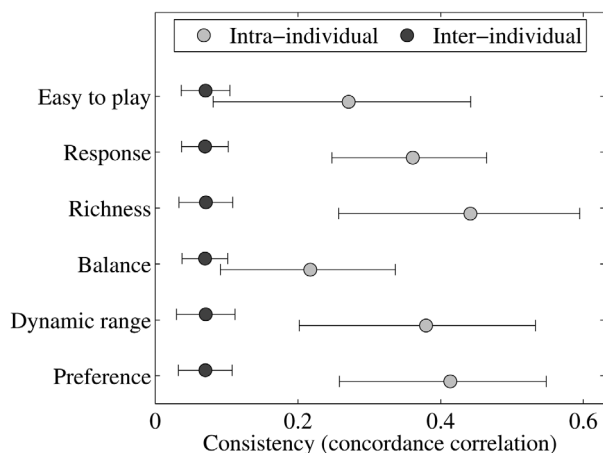

FIG. 4. Experiment 2: Across-participants average intra- and inter-individual consistency scores for each of the attribute-rating scales and preference (error bar = 95% confidence interval of the mean). The preference scores are given at the bottom. See the text for details on averaging of concordance correlations.

same approach described for the analysis of the results of experiment 1. Figure 4 shows the across-participants average of the intra- and inter-individual consistency scores measured for each of the attribute-rating scales. Interestingly, for each of the attribute-rating scales the measures of inter-individual consistency were significantly higher than zero [$t(12) \geq 3.38$, $p \leq 0.006$]. Intra-individual consistency was also significantly higher than zero for all attribute-rating scales [$t(12) \geq 3.08$, $p \leq 0.01$], and was significantly higher than inter-individual consistency [paired samples $t(12) \geq 2.21$, $p \leq 0.047$]. No significant difference emerged between the inter-individual consistency measured for the preference scale, on the one hand, and any of the other attribute scales, on the other [absolute value of paired samples $t(12) \leq 0.18$, $p \geq 0.858$]. Finally, the analysis of intra-individual consistency revealed no significant difference between the preference and any of the attribute scales [absolute value of paired samples $t(12) \leq 1.87$, $p \leq 0.086$], with the exception of a significantly lower level of intra-individual consistency for *balance* than for preference [paired samples $t(12) = -3.18$, $p \leq 0.008$].

We compared the overall measures of intra- and inter-individual consistency collected during experiment 1 with those measured during experiment 2 for the preference rating scale. Intra-individual consistency for the evaluation of preference was significantly higher in experiment 1 than in experiment 2, average value = 0.903 and 0.414, respectively, [independent samples $t(20.17) = 2.25$, $p = 0.036$, unequal variance]. Inter-individual consistency in the evaluation of preference was instead slightly higher in experiment 2 than in experiment 1, average value = 0.071 and 0.017, respectively, although the difference fell short of significance [independent samples $t(30.9) = 1.98$, $p = 0.058$, unequal variance].

2. Influence of participant characteristics

For each of the rating scales, we assessed the association between the participant-specific measures of intra-individual consistency on the one hand, and the self-reported price of the owned violin, the years of violin training, and the weekly hours of violin practice on the other. As for experiment 1, this analysis was carried out by computing the Spearman rank correlation ρ_S between intra-individual consistency scores and participant characteristics. No association was significant [absolute value of $\rho_S \leq 0.402$, $p \geq 0.174$, $df = 11$].[3] Given the small number of amateur as compared to professional violin players who participated in this study ($N = 2$ and 11, respectively), no t-test was carried out to assess the effects of this last participant characteristic on the measures of intra-individual consistency.

3. Preference ranking of the violins

For each participant, and for each of the violins, we then computed a preference score defined as the proportion of times a violin was rated as more preferred than any of the other violins throughout all trials (i.e., we considered only the preference ratings from each trial). The across-participants average preference scores for each violin are reported in Table III and shown in Fig. 5.

4. Relationship between preference and attribute ratings

The remaining analyses assessed the relationship between preference and attribute ratings. All analyses were carried out on the participant-specific attribute ratings along each of the scales, averaged across trials. For each of the participants, we initially estimated a multiple rank-regression model to predict the ranks of the preference ratings based on the ranks of the ratings along the attribute scales. Across participants, attribute ratings predicted a very large amount of variance of the ranked preferences, average $R^2 = 0.901$, SD $= 0.106$. One possible interpretation for this result is that participants used a highly economic response strategy that led them to give strongly correlated ratings along each of the scales irrespective of their actual beliefs about the relationship between violin attributes and preference, thus resulting in a strong association between all scales, preference included—e.g., a hypothetical participant assigning similar ratings to all scales. In order to avoid this possible interpretation, further analyses on the relationship between preference and attribute ratings were carried out based on partial rank correlation coefficients ρ_p.[4] For each participant we computed the partial rank correlation between the preference ratings on the one hand, and the ratings on each of the attribute scales on the other, while controlling for the ratings along the remaining scales. Figure 6 reports the across-participant average of the ρ_p coefficients. Interestingly, preference ratings were significantly associated with two violin attributes: *richness,* average $\rho_p = 0.719$ [$t(11) = 6.05$, $p < 0.001$], and *dynamic range,* average $\rho_p = 0.337$, [$t(11) = 2.59$, $p = 0.024$]. Thus, despite the large amount of inter-individual differences in the evaluation of the different attributes and of preference, participants consistently preferred violins with a richer sound, and with a wider dynamic range. Across participants, none of the other ρ_p coefficients was significant, average absolute $\rho_p \leq 0.084$, [absolute

$t(11) \leq 0.49$, $p \geq 0.636$]. Table III and Fig. 5 also report the across-participants average of the sound richness and dynamic-range width scores for each of the violins. These scores were computed by following the same procedure as for the preference scores (e.g., richness scores = proportion of times a violin is rated as having a richer sound than any of the other violins).

C. Discussion

The results of this experiment showed that experienced violin players are relatively self-consistent when evaluating different violins based on certain characteristics of the instrument as well as in terms of preference. No significant differences were observed between the level of intra-individual consistency in the preference ratings and that in the attribute ratings, with the exception of *balance,* for which self-consistency was significantly lower than that observed for preference. Only two players reported being bothered by the dark sunglasses, whereas no participant reported that the task was difficult overall. Similarly to experiment 1, attempts to associate self-consistency with known (self-recorded) characteristics of the participants were largely inconclusive. Results also confirmed the large inter-individual differences in the preference for violins, while revealing similarly large variations between individual players in rating various violins attributes. The level of inter-individual consistency in each of the attribute-rating scales was not significantly different from that observed in the preference ratings.

Perhaps more importantly, participants were significantly more self-consistent when evaluating preference in experiment 1 than in experiment 2. Many methodological differences between the two experiments could explain this effect. The higher number of trials in experiment 1 (10 rankings of each violin across the two sessions) than in experiment 2 (three ratings of each violin) gave participants a better opportunity to stabilize their response criteria and to accumulate more experience with the evaluated violins. The
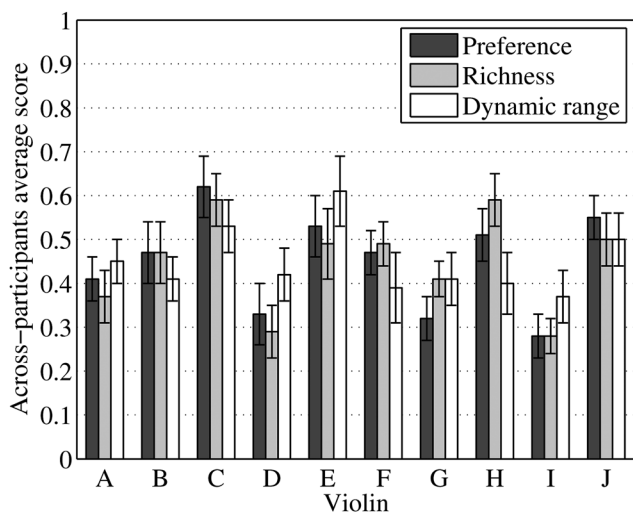

FIG. 5. Experiment 2: Across-participants average of the preference, sound richness and dynamic-range width scores for each violin (error bar = 95% confidence interval of the mean). The violins are ordered by decreasing price. Violins C and I were the most and least preferred, respectively. See the text for details on computing of scores.
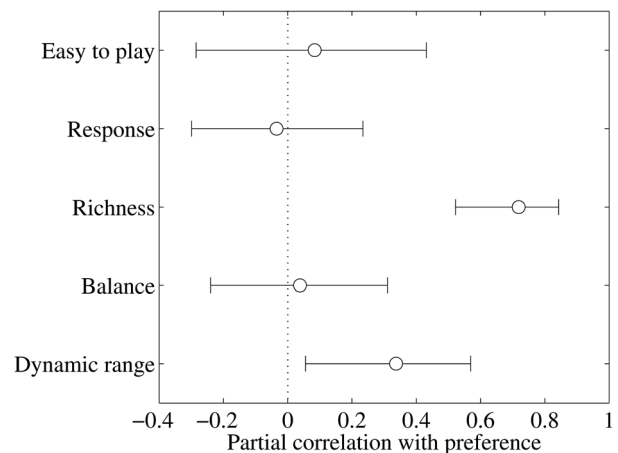

FIG. 6. Experiment 2: Across-participants average partial Spearman rank correlation between violin-preference ratings and ratings along the attribute scales (error bar = 95% confidence interval of the mean). Negative values correspond to anti-consistency (i.e., opposite rankings given on different trials). Only *richness* and *dynamic range* were significantly associated with preference. See the text for details on averaging of correlations.

presence of multiple response scales in experiment 2 but not in experiment 1 did not allow participants in experiment 2 to evaluate preference with the same level of attention as during experiment 1. Finally, due to experimental time constraints, participants in experiment 2 had to rate all criteria, including preference, for a given violin rather than being able to compare the various violins to determine ratings for a criterion.

When evaluating a violin according to specific criteria, players will have their own weightings that define how important each criterion is for them. According to the regression analysis, preference prediction from individual weightings was very high in this experiment, meaning individual players appeared to make their preference judgments by taking into account the various attributes that emerged from the analysis of the verbal data from experiment 1, and using a relatively consistent weighting of these attributes to determine their overall preference for an instrument. A further examination of the association between preference ratings and violin attributes based on measures of partial rank correlation revealed that participants strongly agreed in preferring violins with a rich sound and, to a lesser extent, a wide dynamic range. Combined with the observed low level of inter-individual consistency in both the preference ratings and the ratings on the different attributes, these results show that whereas violinists tend to agree of what particular qualities they look for in an instrument (in this case, sound *richness* and a large *dynamic range*), the perceptual evaluation of the same attributes strongly varies across individuals, thus likely resulting in large inter-individual differences in the preference for violins.

A final consideration is necessary about the interpretation of the large variability in the preference judgments by experienced violinists. Concerning the origin of inter-individual differences in the preference for violins (see Sec. II C), the above observations seem to support, at least in part, the second hypothesis, that different players may follow different perceptual processes to assess different attributes of the violin. On the other hand, there remains the issue of varying playing approaches taken by players to assess different attributes. In this experiment, no playing constraints were imposed on the evaluation process (e.g., specific repertoire). Participants were instead instructed to follow their own strategy with respect to what and how to play. The only way we could discuss this issue further is if we prescribed the musical gestures and/or material that they were allowed to use for the evaluation task. And that still would not address differences in the way people play. Different violinists may use different combinations of gestures when playing, each producing a fundamentally different behavior of the instrument for a certain criterion. For example, player A may use more bow force than player B and thus produce a more "bright" timbre (Schoonderwaldt *et al.*, 2008).

IV. CONCLUSIONS

What is a "good" violin? Most published scientific research on the evaluation of violin qualities has traditionally focused on the physics and mechanics of the instrument and less on the perceptual dimensions related to the player. Indeed, the advent of experimental (e.g., laser-Doppler vibrometry) and computational (e.g., finite element modeling) modal analysis methods in the last decades has provided a comprehensive understanding of the complex acoustical behavior of the violin (e.g., Bissinger and Kuntao, 2000; Roberts, 1997). However, attempts to draw correlations between measured vibrational properties and perceptual judgments have largely been inconclusive. Previous results have demonstrated the need to better understand how violin players perceptually assess different qualities of the instrument. Previous studies have shown that listening tests are not completely indicative of the perceptual processes involved in this context. Indeed, playing-based evaluations are more ecologically valid. Notably, however, no previous study has investigated the extent to which skilled players are consistent at assessing violins and whether there is agreement between violinists to begin with.

Two experiments were carried out based on a carefully controlled playing-based procedure for the perceptual evaluation of violins. We investigated intra-individual consistency and inter-individual agreement in preference judgments by experienced violinists. The results of experiment 1 showed that players are self-consistent when assessing different violins. However, a large amount of inter-individual variability was present in the preference rankings. Overall, known characteristics of the participants (e.g., years of violin training) did not appear to explain self-consistency. The results of experiment 2, wherein preference for the violins was evaluated alongside specific criteria-attributes of the violin, showed that the perception of the same violin attributes widely varied between individual players and corroborated the large inter-individual differences in the preference for the violins observed in experiment 1. Importantly, despite the variability in the evaluation of both preference and violin attributes, an association between preference ratings and ratings on two violin attributes was present. Violinists appeared to strongly agree on their preference for violins with a rich sound and, to a lesser extent, a large dynamic range. As such, what makes a violin good might, to a certain extent, lie in the ears and hands of the performer not because different performers prefer violins with largely different qualities, but because the perceptual evaluation of violin attributes widely considered to be important for a good violin vary across individuals. This important conclusion may explain the limited success of previous studies at quantifying the differences between good and bad violins from vibrational measurements. Further exploration is necessary to better understand how these qualities are perceptually and cognitively evaluated by violinists, and to tease apart the effects of the playing skills of different individuals.

[1]While interesting in itself and certainly worthy of future investigation, the fact that players could easily distinguish poorly maintained Suzuki instruments led us to omit them from consideration because we felt they would skew the overall consistency results. That said, we did make use of a fairly cheap but better maintained violin.

[2]Parametric statistical inferences and averaging of all correlations was carried out on Fisher Z-transformed correlations (Fisher, 1915) in order to attenuate the dependence of the shape of the sampling distribution for the correlation coefficients on the value of the population-average correlation, and in order to minimize biases in the estimation of population-average ρ_c coefficients, which are stronger when averaging is carried out in the raw-correlation space (Silver and Dunlap, 1987). The average ρ_c coefficients and their confidence intervals reported in this manuscript are computed in the Fisher-Z space and transformed back to the raw correlation space by applying the inverse of the Fisher-Z transform (for details, see Thorndike, 2007).

[3]We did observe a significant decrease in the intra-individual consistency for the *response* scale with increasing number of weekly violin-practice hours ($\rho_S = -0.590$, $p = 0.034$, $df = 11$). It should be nonetheless emphasized that this significant result is likely a false positive. Indeed, after a very lenient control of the false-positive rate (Bonferroni-corrected critical p-value, adjusted for the number of participant characteristics $= 0.05/3$), none of the ρ_S coefficients was significant.

[4]The partial correlation $\rho_p(A, B \times C)$ between variables A and B after controlling for variable C is the correlation of the residuals of the regression model that predicts A from C with the residuals of the regression model that predicts B from C. As such, $\rho_p(A, B \times C)$ assesses the association between A and B after eliminating the variance that both A and B share with the controlled variable C. For example, ρ_p (preference, richness × nonpreference, and nonrichness scales) measures the association between ratings along the preference and richness scales after removing the variance that preference and richness ratings share with ratings along the other scales.

Alonso Moral, J., and Jansson, E. V. (**1982**). "Input admittance, eigenmodes and quality of violins," Technical Report No. STL-QPSR 2-3, Department for Speech, Music and Hearing, Royal Institute of Technology, Stockholm, Sweden.

Askenfelt, A. (**1989**). "Measurement of the bowing parameters in violin playing. II: Bow-bridge distance, dynamic range, and limits of bow force," J. Acoust. Soc. Am. **86**, 503–516.

Bissinger, G. (**2008**). "Structural acoustics of good and bad violins," J. Acoust. Soc. Am. **124**, 1764–1773.

Bissinger, G., and Gearhart, F. (**1998**). "A standardized qualitative violin evaluation procedure?," Catgut Acoust. Soc. J. (Series II) **3**, 44–45.

Bissinger, G., and Kuntao, Y. (**2000**). "Automated hammer-impact modal analysis with a scanning laser vibrometer: Working example—a violin," in *Proceedings of the 18th International Modal Analysis Conference*, San Antonio, TX, pp. 943–949.

Caussé, R., Maigret, J. P., Dichtel, C., and Bensoam, J. (**2001**). "Study of violin bow quality," in *Proceedings of the 2001 Internaial Symposium on Musicical Acoustics*, Perugia, Italy.

Fisher, R. (**1915**). "Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population," Biometrika **10**, 507–521.

Fritz, C., Blackwell, A. F., Cross, I., Woodhouse, J., and Moore, B. C. J. (**2012**). "Exploring violin sound quality: Investigating English timbre descriptors and correlating resynthesized acoustical modifications with perceptual properties," J. Acoust. Soc. Am. **131**, 783–794.

Fritz, C., Cross, I., Moore, B. C. J., and Woodhouse, J. (**2007**). "Perceptual thresholds for detecting modifications applied to the acoustical properties of a violin," J. Acoust. Soc. Am. **122**, 3640–3650.

Fritz, C., Woodhouse, J., Cheng, F. P.-H., Cross, I., Blackwell, A. F., and Moore, B. C. J. (**2010**). "Perceptual studies of violin body damping and vibrato," J. Acoust. Soc. Am. **127**, 513–524.

Giordano, B. L., Susini, P., and Bresin, R. (**2012**). "Perceptual evaluation of sound-producing objects," in *Sonic Interaction Design*, edited by K. Franinovic and S. Serafin (MIT Press, Boston, MA).

Inta, R., Smith, J., and Wolfe, J. (**2005**). "Measurement of the effect on violins of ageing and playing," Acoust. Australia **33**, 25–29.

Lin, L. I. (**1989**). "A concordance correlation coefficient to evaluate reproducibility," Biometrics **45**, 255–268.

Loos, U. (**1995**). "Untersuchungen zur Tragfähigkeit von Geigentönen (Studies on the projection of violin tones)," Ph.D. thesis, Department of Media, University of Applied Sciences, Düsseldorf, Germany.

Petiot, J.-F., and Causse, R. (**2007**). "Perceptual differences between cellos: A subjective/objective study," in *Proceedings of the 2007 International Symposium on Musical Acoustics*, Barcelona, Spain.

Roberts, G. W. (**1997**). "Finite element analysis of the violin," in *Research Papers in Violin Acoustics 1975–1993*, edited by C. M. Hutchins and V. Benade (Acoustical Society of America), pp. 575–590.

Schoonderwaldt, E., Guettler, K., and Askenfelt, A. (**2008**). "An empirical investigation of bow-force limits in the Schelleng diagram," Acust. Acust. **94**, 604–622.

Shoukri, M. M. (**2004**). *Measures of Interobserver Agreement* (Chapman and Hall/CRC, Boca Raton, FL).

Silver, N. C., and Dunlap, W. P. (**1987**). "Averaging correlation coefficients: Should Fisher's Z transformation be used?" J. Appl. Psychol. **72**, 146–148.

Thorndike, R. M. (**2007**). "Fisher's Z transformation," in *Encyclopedia of Measurement and Statistics*, edited by N. J. Salkind (Sage, Thousand Oaks, CA), pp. 360–364.

Willgoss, R., and Walker, R. (**2007**). "Discernment of the sound of a violin," in *Proceedings of the 8th World Scientific and Engineering Academy and Society International Conference on Acoustics and Music: Theory and Applications*, Vancouver, Canada, pp. 1–6.

Woodhouse, J. (**1993a**). "On the playability of violins. Part I: Reflection functions," Acustica **78**, 125–136.

Woodhouse, J. (**1993b**). "On the playability of violins. Part II: Minimum bow force and transients," Acustica **78**, 137–153.

Wright, H. (**1996**). "The acoustics and psychoacoustics of the guitar," Ph.D. thesis, Department of Physics and Astronomy, University of Wales, Cardiff, UK.

Zajonc, R. B. (**1968**). "Attitudinal effects of mere exposure," J. Pers. Soc. Psychol. **9**, 1–27.

Saitis *et al.*: Perceptual evaluation of violins