The Journal of the Society for Art and Science, Vol. 21, No. 5, pp. 241–255 (2022)

Useful Feedback in Asynchronous Lessons of Music Performance: A Pilot Study on Oboe Players

Masaki Matsubara[1] **Rina Kagawa**[2]***Takeshi Hirano**[3] **Isao Tsuji**[4,5,6]

1) Faculty of Library, Information and Media Science, University of Tsukuba
2) Faculty of Medicine, University of Tsukuba
3) Graduate School of Information Systems, University of Electro-Communications
4) Nihon University College of Art
5) Kunitachi College of Music
6) Senzoku Gakuen College of Music

masaki (at) slis.tsukuba.ac.jp

Abstract

In the COVID-19 era, the demand for remote and asynchronous lesson of music performance is increasing; however, it is not clear what kind of verbal information should be used. In this study, we collected 239 pieces of textual feedback in Japanese from 12 teachers for 90 performances of the same 10 orchestra studies of oboe performed by nine students. We quantitatively found that the contents of the textual feedback differed most significantly by teacher. Then, we performed multilevel modeling based on hierarchy among teachers to examine usefulness of contents, and found that four types of content contribute usefulness of the textual feedback. The results of a survey of students also supported our analysis, and in addition, suggested that ambiguous statements should be reduced to further improve the usefulness of textual feedback.

*The first two authors equally contributed to this research.

1 Introduction

[1] Playing music instruments has traditionally been taught face-to-face and considered unsuitable for virtual learning environments. However, the COVID-19 pandemic has led to an increase in demand for online music education [3, 4]. In the field of musical performance education, knowledge is conveyed by using both non-verbal information, such as singing melodies and making gestures, and verbal information, such as pointing out mistakes [5, 6, 7]. Verbal information is essential for conveying how the learner's performance sounds, why they make mistakes and how they should practice. In other words, verbal information plays a significant role in the utilization of non-verbal information. An advantage of online music education is that space and time do not necessarily have to be shared, thus allowing for remote and asynchronous teaching. However, the low resolution of online video/audio communication limits the use of non-verbal information, as it is difficult to convey detailed body movements and high-quality sound performances. Therefore, it is expected that the importance of verbal information in music education, especially in the textual feedback of asynchronous online performance education, will increase in the future [8].

However, it is not easy to teach music performance only by words. In our preliminary survey of nine music college students and 100 people who have musical performance experience, most had a good impression of their musical experience, but some were not satisfied with their teacher's instructions. We collected free-text responses about dissatisfaction with the instruction and categorized the results into the following three issues: (1) Inappropriate instruction (e.g., "I would have preferred instruction based on facts," "Lack of concrete advice"); (2) Inconsistent instruction over multiple lessons (e.g., "Completely different or inconsistent attention from lesson to lesson"); and (3) Wording of instruction not related to performance (e.g., "All he/she did was scold without much praise").

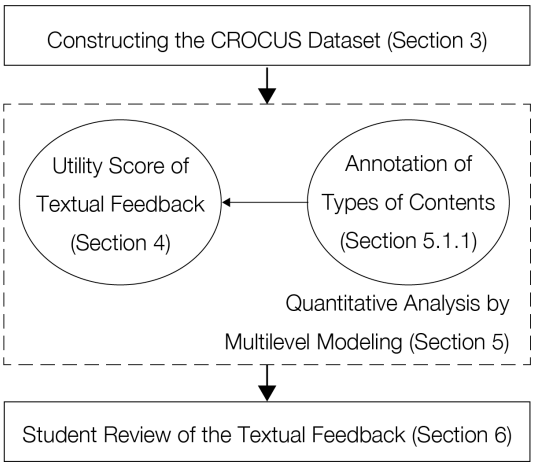


Figure 1: Overview of this manuscripts.

We assume that the reasons for these problems are the lack of teaching protocols in performance instruction and the lack of systematic clarification of what should be verbalized to benefit the learners. At present, however, empirical knowledge of the kinds of instruction that are being given is not widely available, even among students who aspire to become professionals.

Therefore, this study aims to identify what kind of feedback is given, and which of these elements constitute useful feedback.

1.1 Contributions

The overview and contributions of this manuscript are shown as follows (Figure 1).

We constructed an open dataset of musical performance critiques to promote music education for the study of verbal information in performance instruction (Section 3). From this dataset, we quantitatively confirmed that the usefulness of musical performance critiques varies with each feedback, and furthermore, that there is a hierarchy of usefulness among the teachers (Section 4). Therefore, we conducted a multilevel analysis,

[1] A part of this manuscript has been presented at the 15th International Symposium on Computer Music Multidisciplinary Research (CMMR 2021) and at the 23rd International Conference on Asia-Pacific Digital Libraries (ICADL 2021) [1, 2]. The CMMR paper describes the construction of a database of critique documents and the relationship between document structure and utility, while the ICADL paper describes a multilevel model analysis of the relationship between document structure and utility. This manuscript organizes and summarizes the results of the CMMR and ICADL papers, with additional discussion in Section 6, Figure 1 and further results in Table 3 - 5, and revisions to Figure 4, references, and English expressions.

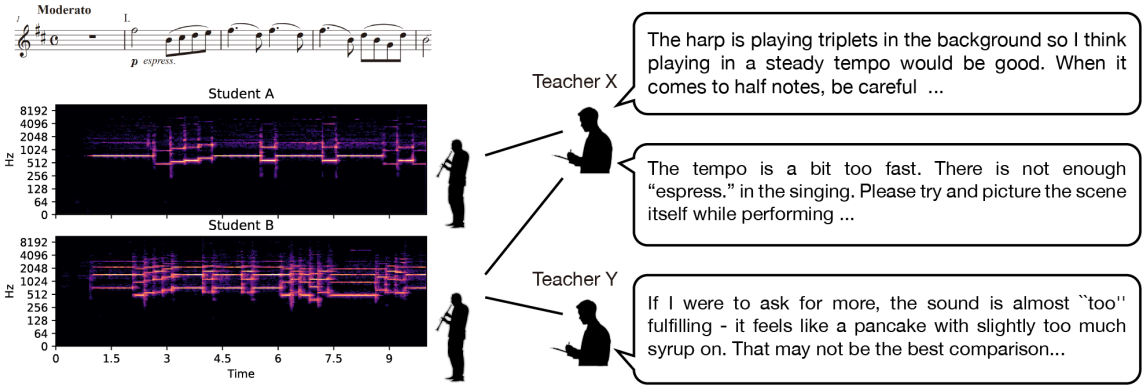


Figure 2: Examples of the CROCUS dataset. Spectrogram of performance recordings of the same piece by two students (left). Textual feedback from teachers X and Y (right). The tempo and the expression vary between students. Teachers express varying points of view.

which allows regression analysis that takes into account the hierarchy of the data, and found that the usefulness was significantly enhanced when four types of content were included in the textual feedback (Section 5). We also conducted a survey of the students of our dataset, which supported our analysis, and also suggested that ambiguous statements should be reduced to further improve the usefulness of textual feedback (Section 6).

2 Related Work

2.1 Music Database for Research

Several public datasets and digital archives have been constructed as music knowledge resources [9]. They adopt various perspectives, including performance recordings data [10], metadata (genre, composer, lyrics, etc. [11, 12, 13]), musical scores (MIDI [14], harmony and cadence [15], and piano notation [16, 17]), information associated with fingering [18] or music analysis [19], other multimodal information [20, 21], emotions [22, 23], listening history [24], and performers' interpretations [25, 26, 27, 28]. To the best of our knowledge, none have focused on human cognition, such as the experiences of playing or listening to a piece of music.

2.2 Effects of Teaching Behavior on Musical Performance Education

The relationship between teaching behavior and musical performance education has been widely studied within the field of music education. Prior studies have focused upon comparison; e.g. comparison of teacher levels [29], analysis of time allocation [30], comparison [31] and categorization [32, 33] of verbal and non-verbal information, and teacher-student interaction [34]. These studies all depended upon the transcription of speech in interactive instruction. Our study focused on textual feedback, which is more applicable to asynchronous education.

One study compared verbal and non-verbal instruction [35], and another study summarized the evaluation of the usefulness [36]. Both were based on five or fewer performances. In contrast, we conducted a large-scale experiment and successfully clarified the relationship between verbal information and utility.

3 Constructing the CROCUS Dataset

We first construct *CROCUS* (CRitique dOCUmentS of musical performance) dataset[2] by collecting the performance recordings and textual feedback (Figure 2).

[2]Available on https://doi.org/10.5281/zenodo.4748243

Table 1: List of pieces of the CROCUS dataset

ID	Composer	Piece
p01	L. v. Beethoven	Symphony No. 3 in E flat Major "Eroica," Op. 55
p02	G. A. Rossini	"La Scala di seta" Overture
p03	F. Schubert	Symphony No. 8 in B Minor D.759 "Unfinished"
p04	J. Brahms	Violin Concerto in D Major, Op. 77
p05	P. I. Tchaikovsky	Symphony No. 4 in F minor, Op. 36
p06	P. I. Tchaikovsky	"Swan Lake," Ballet Suite, Op.20a
p07	N. Rimsky-Korsakov	"Scheherazade," Symphonic Suite, Op. 35
p08	R. Strauss	"Don Juan," Symphonic Poem, Op. 20
p09	M. Ravel	Le Tombeau de Couperin I.Prelude
p10	S. Prokofiev	"Peter and the Wolf," Symphonic Tale, Op. 67

3.1 Methods

3.1.1 Recording

A total of 90 performances were recorded (10 orchestral studies performed by nine music college students majoring in oboe). From the oboe orchestra study[3], we selected the 10 pieces shown in Table 1, considering a balance of difficulty, style, form, and era. In this study, each student played in an environment with limited reverberation and noise, at home, about one meter away from the recording device (Roland R-07). Tuning and recording level were adjusted at the beginning of the recording.

3.1.2 Textual feedback for each performance recording

Twelve teachers were participated. Each teacher wrote one piece of textual feedback, assuming usual lessons, for each performance recording. Each teacher wrote 20 textual feedback. The 20 performances were selected in a counterbalanced manner by following constraints; each teacher reviewed two performances for each piece, and each student was reviewed from all the teachers throughout the ten performances. Audio files of performance recording were sent to each teacher, along with an introduction: "Please write textual feedback for each recording assuming the usual lessons." They listened to each recording and either wrote or typed their feedback.

How to collect the performance recordings and textual feedback described above was decided based on the fact: due to the influence of COVID-19, music students at Japanese College of Music recorded performances at home and sent the recordings to the teachers. After the teachers described the instruction comments for the recording, they returned the instruction texts to the students.

Both students and teachers were informed that recording and collecting textual feedback in this study was aimed to improve music performance education, but not told more complicated methods or objects of our experiments. The students do not know who the teachers are, and they do not know the correspondence between the review and the teacher.

3.2 Results

A total of 239 performance critiques were collected, as one teacher missed to write one performance critique.

4 The Utility Score of Each Piece of Feedback

4.1 Methods

We examined the perceived utility by performers who read the collected critiques.

[3]Oboe orchestra study is the collection of the pieces of orchestra for oboe, which is widely used all over the world for examinations in the music colleges or audition for the professional orchestra.

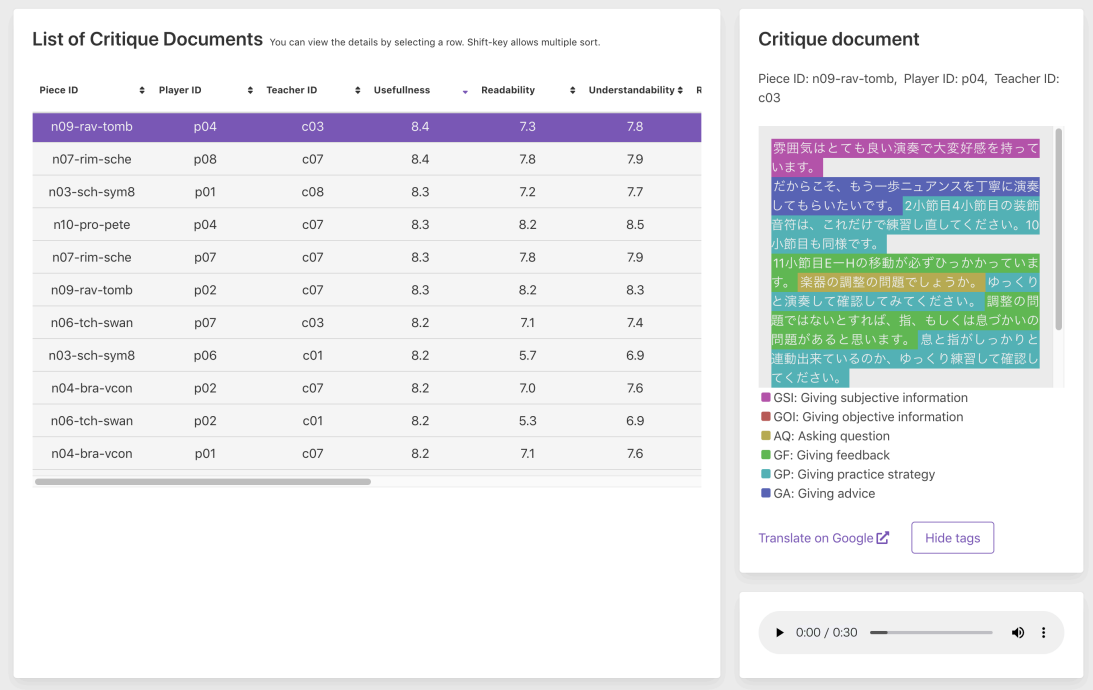


Figure 3: The screenshot of the demonstration page of CROCUS.

A total of 200 people who have musical experience answered the question ("Do you think that this document is useful for future performances?") on an 11-point Likert (10: useful – 0: useless)[4]. Participants responded to 25 randomly selected textual feedback. In this manuscript, the score on the 11-point Likert scale is called the utility score.

4.2 Results

The utility score for each piece of textual feedback is shared on the demonstration page (Figure 3).[5] The textual feedback with the highest average utility score and that with the lowest average score are shown below.

The highest rated critique (utility score: 8.41 ± 1.44)
I feel that this performance is very good, and it leaves a very favorable impression. Because of this, I would like you to be a little more careful in regards to the nuances of the performance. Please practice the grace notes in bars 2 and 4 again by themselves. The same for bar 10. There is always a mistake in the E-H transition in bar 11. Perhaps it is a problem with the tuning of the instrument. Please perform this part slowly and check carefully. If it is not a tuning problem, then I believe it is a fingering or breathing problem. Please practice carefully and check if the breathing and fingering are both coordinated properly. In the second half, there is tenuto on the high E and D notes. Please endeavour to perform each note carefully with nuance.

The lowest rated critique (utility score: 4.63 ± 2.61)
The melodies are performed beautifully and vibrantly, almost as if I could hear an orchestra performing. The phrasings are well expressed for the piece, and it was lovely.

[4]The number of Likert scale points to use is still controversial. We referred to research [37] showing that more scale points led to reduced skewness and normal distributions, and we adopted an 11-point scale.
[5]https://masaki-cb.github.io/crocus/

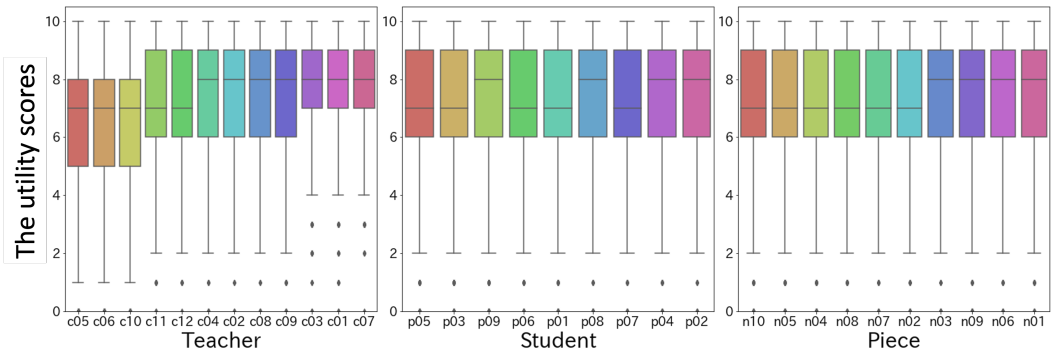


Figure 4: Average utility scores based on teacher, student, and piece. **Remark:** *Average utility scores differed depending on the teachers.*

Figure 4 shows the average utility scores for each teacher, student, and piece. This result implies that the usefulness of the critiques differed more by the teacher than by the piece or student.

For teachers, the intraclass correlation coefficient (ICC) was 0.45, and the design effect (DE)[1] was 9.43. For players or pieces, ICCs were 0.0, and DEs were 1.0. Therefore, the utility scores showed hierarchy among teachers.

5 Quantitative Analysis of Content that Contributes to Usefulness of Instruction

The analysis so far has revealed that the utility scores of instructional documents is stratified by teachers. Therefore, we quantitatively examined how the contents of the description differ depending on the teachers and which contents actually contribute to the usefulness.

5.1 Methods

5.1.1 Annotation of Types of Contents

As this study considers asynchronous instruction with textual feedback, we adopted and adapted the types in Simones' definition [32], Carlin, 1997 [38], and Zhukov, 2004 [39] as shown in the Table 2[6], which is a classification created by reviewing qualitative points that should be included in music performance instruction in the field of music education. One of these six types of contents was annotated to each sentence; sentence breaks were considered to be periods or exclamation marks. When it was judged that one sentence consisted of descriptions correlating to multiple kinds of types, commas were used to separate the relevant sections.

Each two annotators annotated all 239 documents. If the annotations did not match, the final annotation was decided through discussion. The Cohen's Kappa coefficient was 0.96.

For the annotation results, the percentages of documents containing GSI, GOI, AQ, GF, GP, and GA were 47.28%, 54.81%, 3.34%, 39.33%, 22.18%, and 93.72%, respectively. The average (and standard deviation) of the number of sentences annotated as each type, per document, was 0.70 (0.90), 0.85 (1.00), 0.03 (0.18), 0.61 (0.88), 0.33 (0.70), and 3.33 (2.50), respectively.

For the hierarchy among teachers in terms of the number of sentences annotated as each type, ICC and DE in the order of GSI, GOI, AQ, GF, GP, and GA were 0.24 and 5.54, 0.04 and 1.76,

[1] DE is a criterion that takes into account both the average number of data in the group and ICC. $DE = 1+(k^*-1)ICC$. k^* means the average number of data of the group. An ICC was over 0.05 or a DE of over two suggested that the data were hierarchical.

[6] Types of "Demonstrating," "Modelling," and "Listening/Observing" were omitted because these actions are not observed in textual critique. In the Simones' definition, Giving Information is one category, but authors divided this in Giving Subjective Information and Giving Objective Information.

Table 2: Types of verbal information used in this study

Types	Definition	Example of sentence
Giving Subjective Information (GSI)	Teacher providing general and/or specific conceptual information based on teacher's subjectivity.	*It is a very light and springy performance.*
Giving Objective Information (GOI)	Teacher providing general and/or specific conceptual information based on objectively referable events or concepts.	*Tempo is late in the second bar.*
Asking Question (AQ)	Enquiring.	*Is there a problem with the tuning of the instrument?*
Giving Feedback (GF)	Teacher evaluation of a student's applied and/or conceptual knowledge.	*The pitch unconsciously moves during a vibrato.*
Giving Practice (GP)	Providing suggestions of ways to practice a particular passage or discussing a practicing schedule.	*Please practice this phrase using the metronome.*
Giving Advice (GA)	Giving a specific opinion or recommendation without demonstration or modelling to guide the student's action towards the achievement of certain specific musical aims.	*The first bar should have no crescendo.*

0.11 and 3.14, 0.15 and 3.92, 0.11 and 3.04, and 0.52 and 10.78, respectively. This confirms that there is a hierarchy among the teachers in terms of the number of sentences of all types of contents except GOI. For the hierarchy among teachers in terms of the presence or absence of sentences annotated as each type, ICC and DE in the order of GSI, GOI, AQ, GF, GP, and GA were 0.15 and 3.80, 0.03 and 1.48, 0.11 and 3.14, 0.09 and 2.76, 0.11 and 3.08, and 0.09 and 2.70, respectively. This confirms that there is a hierarchy among the teachers in terms of the presence or absence of sentences of all types of contents except GOI, GF, or GA.

5.1.2 Multilevel Modeling

Multilevel modeling was conducted to quantitatively analyze the effect of the presence or absence of sentences annotated as each type of contents or the number of sentences annotated as each type of contents on the utility score. Multilevel modeling enables analysis assuming that the behavior of individual data changes depending on the hierarchy of data. In other words, in this study, not only the change in utility scores among documents but also the influence of the teachers could be analyzed. We first tested the hierarchy of the characteristics of documents and then devised four models for analysis. R 4.1.0, brms 2.15.0, lme4 1.1–27, and lattice 0.20–38 were used.

Based on the observed hierarchy, we devised the following four models focusing on the presence or absence of each of the six types of contents and the number of sentences of each type.

Model I: The utility scores of the documents is affected by the presence or absence of each type.

Model II: The utility scores of the documents is affected by the number of sentences of each type.

Model III: The utility scores of the documents is affected by the presence or absence of each type and varies depending on teachers.

Table 3: Effect of the presence or absence of sentences of each type and teacher on document utility scores (Model I). GSI_{yn}, ... , GA_{yn} in this table means the presence or absence of sentences of each corresponding type.

Population-Level Effects	Estimate	Est.Error	lower-95% CI	upper-95% CI	$\hat{R}$
Intercept	6.16	0.19	5.78	6.54	1.00
GSI_{yn}	0.00	0.08	−0.16	0.17	1.00
GOI_{yn}	**0.35**	0.08	**0.20**	**0.50**	1.00
AQ_{yn}	−0.44	0.22	−0.87	−0.01	1.00
GF_{yn}	0.13	0.08	−0.03	0.29	1.00
GP_{yn}	**0.52**	0.10	**0.32**	**0.72**	1.00
GA_{yn}	**0.81**	0.17	**0.48**	**1.15**	1.00
Family Specific Parameters					
Sigma	0.61	0.03	0.56	0.66	1.00

Table 4: Effect of the number of sentences of each type and teacher on document utility scores (Model II). GSI, ... , GA in this table means the number of sentences of each corresponding type.

Population-Level Effects	Estimate	Est.Error	lower-95% CI	upper-95% CI	$\hat{R}$
Intercept	6.45	0.09	6.28	6.62	1.00
GSI	0.03	0.04	−0.05	0.11	1.00
GOI	**0.17**	0.03	**0.10**	**0.24**	1.00
AQ	−0.29	0.20	−0.67	0.10	1.00
GF	**0.16**	0.04	**0.08**	**0.24**	1.00
GP	**0.29**	0.05	**0.19**	**0.39**	1.00
GA	**0.14**	0.01	**0.11**	**0.17**	1.00
Family Specific Parameters					
Sigma	0.53	0.02	0.48	0.58	1.00

Model IV: The utility scores of the documents is affected by the number of sentences of each type and varies by teachers.

Let α be intercept, k be a content category, $\beta_k (k = 1, \ldots, 6)$ be the coefficient of n_{ki}, and n_{ki} be the number of descriptions for each type. In the i-th document of the j-th participants, the utility scores of the k-th content U_{ij} is designated as follows:

$$U_{ij} = \alpha + \sum_{k=1}^{6} \beta_k n_{jk} + \sum_{k=1}^{6} \eta_k^{(z_{ijk})} + \sum_{k=1}^{6} \gamma_k^{(z_{ijk})} n_{ik} + e_{ij}$$

Here, z_{ijk} indicates each teacher who wrote the i-th document. $\beta_k^{(z_{ijk})}$ is the random effect of the presence of unknown words on the intercept for the k-th content category of the i-th document. $\gamma_k^{(z_{ijk})}$ is the random effect of the presence of unknown words on the coefficient for n_{ik}.

The model parameters were fitted with four Markov chain Monte Carlo chains with 2,000 iterations and 1,000 burn-in samples with a thinning parameter of one. Non-informative priors were used for all estimations. Specifically, we used $\beta_k \sim N(0, 100)$, $\alpha \sim StudentT(3, 0, 2.5)$, and $\sigma_e \sim StudentT(3, 0, 2.5)$ as the prior distributions of the fixed effects, $StudentT(3, 0, 2.5)$ as the prior distribution of SD of random effects, and $LKJCholesky(1)$ as the prior distribution of the correlation matrix between $\gamma_k^{(g)}$ and $\eta_k^{(g)}$ for $k \in \{1, \ldots, 6\}$ and $g \in \{1, \ldots, 12\}$. The models were compared based on the widely applicable information criterion (WAIC). A smaller WAIC corresponds to a better model.

5.2 Results

As fitting indices, the WAIC values for models I–IV were 449.6, 381.7, 320.0, and 292.1, respectively. All $\hat{R}$ were 1.01 or less. These results indicate that model IV was the best model; that is, the number of sentences of all types affects the utility scores, and these influences are affected by

Table 5: Effect of the presence or absence of sentences of each type and teacher on document utility scores (Model III). GSI_{yn}, ... , GA_{yn} in this table means the presence or absence of sentences of each corresponding type.

Population-Level Effects	Estimate	Est.Error	lower-95% CI	upper-95% CI	$\hat{R}$
Intercept	6.43	0.40	5.64	7.28	1.00
GSI_{yn}	0.09	0.14	−0.18	0.37	1.00
GOI_{yn}	**0.25**	0.08	**0.10**	**0.40**	1.00
AQ_{yn}	0.57	0.81	−1.21	2.15	1.00
GF_{yn}	0.12	0.10	−0.07	0.30	1.00
GP_{yn}	0.32	0.18	−0.03	0.68	1.00
GA_{yn}	0.56	0.04	−0.03	1.17	1.00
Group-Level Effects					
sd(Intercept)	0.30	0.25	0.01	0.92	1.00
$\mathrm{sd}(\mathrm{GSI}_{yn})$	0.38	0.15	0.12	0.72	1.00
$\mathrm{cor}(\mathrm{Intercept},\mathrm{GSI}_{yn})$	−0.19	0.57	−0.97	0.92	1.01
sd(Intercept)	0.32	0.26	0.01	0.96	1.00
$\mathrm{sd}(\mathrm{GOI}_{yn})$	0.11	0.09	0.00	0.32	1.00
$\mathrm{cor}(\mathrm{Intercept},\mathrm{GOI}_{yn})$	−0.18	0.57	−0.97	0.92	1.00
sd(Intercept)	0.30	0.25	0.01	0.97	1.00
$\mathrm{sd}(\mathrm{AQ}_{yn})$	1.37	0.84	0.41	3.49	1.01
$\mathrm{cor}(\mathrm{Intercept},\mathrm{AQ}_{yn})$	−0.04	0.57	−0.95	0.94	1.00
sd(Intercept)	0.30	0.25	0.01	0.90	1.00
$\mathrm{sd}(\mathrm{GF}_{yn})$	0.21	0.11	0.02	0.46	1.01
$\mathrm{cor}(\mathrm{Intercept},\mathrm{GF}_{yn})$	−0.19	0.57	−0.98	0.93	1.00
sd(Intercept)	0.32	0.25	0.01	0.94	1.00
$\mathrm{sd}(\mathrm{GP}_{yn})$	0.43	0.22	0.08	0.94	1.00
$\mathrm{cor}(\mathrm{Intercept},\mathrm{GP}_{yn})$	−0.22	0.56	−0.98	0.91	1.00
sd(Intercept)	0.45	0.35	0.02	1.33	1.00
$\mathrm{sd}(\mathrm{GA}_{yn})$	0.62	0.30	0.16	1.36	1.00
$\mathrm{cor}(\mathrm{Intercept},\mathrm{GA}_{yn})$	−0.44	0.53	−0.99	0.83	1.01
Family Specific Parameters					
Sigma	0.42	0.02	0.38	0.48	1.00

the teachers in all types. From the values of ICC and DE for the six types of contents shown in Section 5.1.1, it was inferred that it was reasonable to use a multilevel model (Model IV) that assumed hierarchy in terms of the effect of the number of sentences meaning the six types of contents on the utility score.

The statistics of Model I - Model IV were shown in Table 3, Table 4, Table 5, and Table 6. We compared Model I-IV based on Table 3-6. Here, we compared Models I and III, which discuss the impact of the presence or absence of sentences meaning the six types of contents on the utility scores, and Models II and IV discussing the impact of the number of sentences meaning the six types of contents on the utility scores. Then, the standard deviation of the intercept and slope of the random effect was significant both when comparing Model I and III (Table 3 and Table 5) and when comparing Model II and IV (Table 4 and Table 6). In other words, it can be concluded that random effects should be considered for both the intercept and the slope. These results also showed model IV was the best model.

Table 6 shows the effect of number of sentences of each type and teacher on utility scores. GOI ($\beta_2 = 0.13, 95\%CI[0.05 - 0.20]$), GF ($\beta_4 = 0.13, 95\%CI[0.04 - 0.23]$), GP ($\beta_5 = 0.27, 95\%CI[0.09 - 0.46]$), and GA ($\beta_6 = 0.15, 95\%CI[0.07 - 0.22]$) had positive lower-95% Credible Interval (CI), among which GP showed the highest estimate score based on Model IV.

The results suggest that the more sentences of GOI, GF, GP, and GA significantly increased

Table 6: Effect of the number of sentences of each type and teacher on document utility scores. GSI, ... , GA in this table means the number of sentences of each corresponding type. **Note:** GOI, GF, GP, and GA had positive lower-95% Credible Interval (CI), among which GP showed the highest estimate score. This indicates that utility scores will increase if the number of the four types is increased and that GP is the most effective type for instruction. **Remark:** *The more sentences of GOI, GF, GP, and GA significantly increased the utility scores, and GP had the highest utility scores among all the models.*

Population-Level Effects	Estimate	Est.Error	lower-95% CI	upper-95% CI	$\hat{R}$
Intercept	6.55	0.24	6.04	7.01	1.00
GSI	0.08	0.08	−0.08	0.23	1.00
GOI	**0.13**	0.04	**0.05**	**0.20**	1.00
AQ	0.54	0.92	−1.31	2.39	1.00
GF	**0.13**	0.05	**0.04**	**0.23**	1.00
GP	**0.27**	0.09	**0.09**	**0.46**	1.00
GA	**0.15**	0.04	**0.07**	**0.22**	1.00
Group-Level Effects					
sd(Intercept)	0.20	0.17	0.01	0.62	1.00
sd(GSI)	0.22	0.08	0.08	0.40	1.00
cor(Intercept,GSI)	−0.21	0.56	−0.97	0.91	1.01
sd(Intercept)	0.21	0.18	0.01	0.65	1.00
sd(GOI)	0.05	0.05	0.00	0.18	1.00
cor(Intercept,GOI)	−0.15	0.58	−0.97	0.93	1.00
sd(Intercept)	0.19	0.17	0.01	0.63	1.00
sd(AQ)	1.47	0.86	0.50	3.73	1.01
cor(Intercept,AQ)	−0.01	0.57	−0.94	0.95	1.00
sd(Intercept)	0.21	0.18	0.01	0.67	1.00
sd(GF)	0.08	0.07	0.00	0.25	1.01
cor(Intercept,GF)	−0.13	0.57	−0.97	0.92	1.00
sd(Intercept)	0.22	0.19	0.01	0.72	1.00
sd(GP)	0.21	0.11	0.03	0.48	1.00
cor(Intercept,GP)	−0.24	0.56	−0.97	0.89	1.00
sd(Intercept)	0.44	0.25	0.03	1.03	1.00
sd(GA)	0.10	0.04	0.03	0.19	1.00
cor(Intercept,GA)	−0.67	0.42	−1.00	0.62	1.01
Family Specific Parameters					
Sigma	0.40	0.02	0.36	0.45	1.00

the utility scores, and GP had the highest utility scores among all the models.

6 Student Review of the Textual Feedback

In order to determine whether the results obtained in the previous section can be applied to the students themselves, and to examine points that should be improved in lesson of music performance in addition to the content, an additional survey was conducted for the students who provided performance recordings to CROCUS.

6.1 Methods

We asked them to write freely about their dissatisfaction with each feedback. If there was nothing in particular that they were dissatisfied with, they wrote "nothing".

For each response, one of the authors performed the coding to determine what was being described. The six contents of Table 2, the 12 elements[7] listed as quality characteristics in

[7]The full list of these elements is Necessary, Appropriate, Unambiguous, Complete, Singular, Feasible, Verifiable, Correct, Conforming, Consistent, Comprehensive, Able to be Validated.

Table 7: Dissatisfaction with textual feedback in students response

ID	Example of dissatisfaction
GSI	*"I thought it would be nice to have some feedback on the performance itself, just briefly."*
GOI	*"I would appreciate it the teacher you could tell me what a natural vibrato and tempo sounds like."*
GP	*"I wanted to know how I could practice to get better."*
GA	*"I would like to know how to play as if I were playing in an orchestra."*
Complete	*"The description was short and did not provide any useful points other than the length of H."*
Unambiguous	*"I was hoping the teacher could give me an example of an articulation that the teacher found strange."*
Dissatisfaction toward the mentioned point	*"It is mentioned that the performance is like an etude, but I felt that this expression is a little inappropriate because there are various types of etudes, such as plain and singing."*
Wording	*"I wish the teacher had chosen his/her words more carefully."*

ISO/IEC/IEEE 29148: 2018[8], which is known as the global standard for requirements in the field of software engineering requirements, or some other element were used for this coding. If more than one element was determined to be applicable to the description, multiple elements were coded.

6.2 Results

In total, 192 feedback responses were collected, and dissatisfaction was mentioned in 29 responses (Table 7).

The results of coding were as follows: GSI, one description; GOI, five descriptions; GA, 12 descriptions; GP, 14 descriptions; Complete, two descriptions; Unambiguous, two descriptions; Dissatisfaction toward mentioned point, one description; and wording, two descriptions. The statement "How should I do it?" was counted as both GA and GP because it could be interpreted as either a lack of GA or a lack of GP.

Based on the results of the utility scores given by amateur musicians with performance experience, it was confirmed that the three elements other than GF (GOI, GP, and GA) all of which were found to significantly increase the utility scores, were also required to teach music performance to music college students. This suggests that the current content of music performance teaching is not sufficient, regardless of the level of performance experience. In the future, it will be necessary to create an index to evaluate the appropriateness of wording and the degree of unambiguity.

7 Discussion

We confirmed that contents of the lessons of music performance differed more by the teacher than by the piece or the student. The number of descriptions of GA, GF, GOI, and GP significantly improved the usefulness of feedbacks.

Although it has been pointed out in previous studies [40, 41] that teachers do not always agree on the evaluation of music performance, to the best of our knowledge, there is no research that has pointed out that there is a hierarchy among teachers in terms of the ratings given by players to the content of their feedback.

7.1 Generalizability and Applicability

The findings revealed for asynchronous lessons in this study can be applied to face-to-face lessons. However, since the discussion is limited to the oboe, we would like to broaden generalizability in the future by examining additional instruments and music genres. Furthermore, students in this

[8] https://www.iso.org/obp/ui/#iso:std:iso-iec-ieee:29148:ed-2:v1:en

experiment are limited to music college students. In the future, we would like to discuss the difference in results due to differences in student levels.

In particular, accumulated textual feedback from lessons has the potential for knowledge transfer and reuse; for example, students can use them for their practice, other students can also use them as references, and teachers can use them to improve their teaching methods. Although previous studies have focused on the transcription of speech in interactive instruction [29, 30, 31, 32, 33, 34, 35, 36], only a few have investigated textual documents in this context.

7.2 Future Works

The interaction between factors was not considered in the current study. For future work, we would like to examine the interaction between types (e.g., when sentences of AG are numerous and the influence of GSI is small and vice versa).

It is also necessary to further investigate the effect of the player's knowledge and the relationship between the player and the teacher on usefulness (e.g., whether a relationship of trust has been established).

There is a need for additional support for teachers who must write textual feedback. To address this problem, future works should consider how to express sentences of the contents of the lesson of music performance and should develop an authoring tool.

8 Conclusion

We published a CROCUS dataset as a starting point for investigating the utility of contents in textual feedback provided as part of musical instrument instruction. This dataset clarified that the contents of the music performance education varied most significantly by teacher. Based on multilevel modeling, we quantitatively found that the number of sentences of six types of contents tended to improve the usefulness of the performance instruction document. Furthermore, the larger the number of sentences of GA, GF, GOI, and GP, the more significant was the increase in the usefulness of the documents. The effect was different depending on the teacher. We found that it will be necessary to improve the quality of performance instruction in order to ensure that the performer learns everything they want to know and ambiguity is avoided. In the future, we would like to discuss the arrangement of documents, determine their format, and consider the development of educational programs and writing support technologies so that teachers can make these sentences.

Acknowledgment

This study was partially supported by JST-Mirai Program Grant Number JPMJMI19G8, and JSPS KAKENHI Grant Number JP19K19347 and JP21H03552.

Ethics

This research was approved by the Ethics Committees of Faculty of Library Information and Media Science University of Tsukuba (permission number:20-22), Kunitachi College of Music (permission number: 2007), and Senzoku Gakuen College of Music. Opt-in informed consent was obtained from performers and teachers who engaged in the experiments. Gratuities were paid for participants in accordance with the regulations of the University of Tsukuba.

References

[1] Masaki Matsubara, Rina Kagawa, Takeshi Hirano, and Isao Tsuji. Crocus: Dataset of musical performance critiques: Relationship between critique content and its utility. In *CMMR*, 2021.

[2] Masaki Matsubara, Rina Kagawa, Takeshi Hirano, and Isao Tsuji. Analysis of the usefulness of critique documents on musical performance: Toward a better instructional document format. In Hao-Ren Ke, Chei Sian Lee, and Kazunari Sugiyama, editors, *Towards Open and Trustworthy Digital Soci-*

eties, pages 344–353. Springer International Publishing, 2021.

[3] Jonathan G Bayley and Janice Waldron. "it's never too late": Adult students and music learning in one online and offline convergent community music school. *Int. J. Music. Educ.*, 38(1):36–51, 2020.

[4] Phillip M Hash. Remote learning in school bands during the covid-19 shutdown. *J. Res. Music. Educ.*, 68(4):381–397, 2021.

[5] Phyllis E Dorman. A review of research on observational systems in the analysis of music teaching. *Bull. Counc. Res. Music. Educ.*, pages 35–44, 1978.

[6] Hildegard Froehlich. Measurement dependability in the systematic observation of music instruction: A review, some questions, and possibilities for a (new?) approach. *Psychomusicology*, 14(1-2):182, 1995.

[7] Andreas C Lehmann, John A Sloboda, Robert Henley Woody, Robert H Woody, et al. *Psychology for musicians: Understanding and acquiring the skills*. Oxford University Press, 2007.

[8] Keith Dye. Student and instructor behaviors in online music lessons: An exploratory study. *Int. J. Music. Educ.*, 34(2):161–170, 2016.

[9] Justin Salamon. What's broken in music informatics research? three uncomfortable statements. In *36th International Conference on Machine Learning (ICML), Workshop on Machine Learning for Music Discovery. Long Beach, CA, USA*, 2019.

[10] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. Rwc music database: Popular, classical and jazz music databases. In *ISMIR*, pages 287–288, 2002.

[11] Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. Rwc music database: Music genre database and musical instrument sound database. In *ISMIR*, pages 229–230, 2003.

[12] Carlos Nascimento Silla Jr, Alessandro L Koerich, and Celso AA Kaestner. The latin music database. In *ISMIR*, pages 451–456, 2008.

[13] Bob L. Sturm. An analysis of the gtzan music genre dataset. In *ACM Workshop MIRUM*, MIRUM '12, pages 7–12, 2012.

[14] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *ICLR*, 2019.

[15] Johannes Hentschel, Markus Neuwirth, and Martin Rohrmeier. The annotated mozart sonatas: Score, harmony, and cadence. *Transactions of the International Society for Music Information Retrieval*, 4(1), 2021.

[16] Francesco Foscarin, Andrew McLeod, Philippe Rigaux, Florent Jacquemard, and Masahiko Sakai. ASAP: a dataset of aligned scores and performances for piano transcription. In *ISMIR*, pages 534–541, 2020.

[17] Ziyu Wang, Ke Chen, Junyan Jiang, Yiyi Zhang, Maoran Xu, Shuqi Dai, Guxian Bin, and Gus Xia. Pop909: A pop-song dataset for music arrangement generation. In *ISMIR*, 2020.

[18] Eita Nakamura, Yasuyuki Saito, and Kazuyoshi Yoshii. Statistical learning and estimation of piano fingering. *Information Sciences*, 517:68–85, 2020.

[19] Masatoshi Hamanaka, Keiji Hirata, and Satoshi Tojo. Gttm database and manual time-span tree generation tool. In *SMC*, pages 462–467, 2018.

[20] Bochen Li, Xinzhao Liu, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma. Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. *IEEE Tran. Multimedia*, 21(2):522–535, 2018.

[21] Christof Weiß, Frank Zalkow, Vlora Arifi-Müller, Meinard Müller, Hendrik Vincent Koops, Anja Volk, and Harald G Grohganz. Schubert winterreise dataset: A multimodal scenario for music analysis. *J. Comp. Cult. Herit.*, 14(2):1–18, 2021.

[22] Yu-An Chen, Yi-Hsuan Yang, Ju-Chiang Wang, and Homer Chen. The amg1608 dataset for music emotion recognition. In *ICASSP*, pages 693–697, 2015.

[23] Kejun Zhang, Hui Zhang, Simeng Li, Changyuan Yang, and Lingyun Sun. The pmemo dataset for music emotion recognition. In *ICMR*, pages 135–142, 2018.

[24] Markus Schedl. The lfm-1b dataset for music retrieval and recommendation. In *ICMR*, pages 103–110, 2016.

[25] Mitsuyo Hashida, Toshie Matsui, and Haruhiro Katayose. A new music database describing deviation information of performance expressions. In *ISMIR*, pages 489–494, 2008.

[26] Mitsuyo Hashida, Eita Nakamura, and Haruhiro Katayose. Constructing pedb 2nd edition: a music performance database with phrase information. In *SMC*, pages 359–364, 2017.

[27] Rolando Miragaia, Gustavo Reis, Francisco Fernandéz de Vega, and Francisco Chávez. Multi pitch estimation of piano music using cartesian genetic programming with spectral harmonic mask. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1800–1807. IEEE, 2020.

[28] Craig Stuart Sapp. Comparative analysis of multiple musical performances. In *ISMIR*, pages 497–500, 2007.

[29] Thomas W Goolsby. Verbal instruction in instrumental rehearsals: A comparison of three career levels and preservice teachers. *J. Res. Music. Educ.*, 45(1):21–40, 1997.

[30] Mary Ellen Cavitt. A descriptive analysis of error correction in instrumental music rehearsals. *J. Res. Music. Educ.*, 51(3):218–230, 2003.

[31] Jennifer A Whitaker. High school band students' and directors' perceptions of verbal and nonverbal teaching behaviors. *J. Res. Music. Educ.*, 59(3):290–309, 2011.

[32] Lilian Lima Simones, Matthew Rodger, and Franziska Schroeder. Communicating musical knowledge through gesture: Piano teachers' gestural behaviours across different levels of student proficiency. *Psychology of Music*, 43(5):723–735, 2015.

[33] Lilian Simones, Franziska Schroeder, and Matthew Rodger. Categorizations of physical gesture in piano teaching: A preliminary enquiry. *Psychology of Music*, 43(1):103–121, 2015.

[34] Robert A Duke and Amy L Simmons. The nature of expertise: Narrative descriptions of 19 common elements observed in the lessons of three renowned artist-teachers. *Bull. Counc. Res. Music. Educ.*, pages 7–19, 2006.

[35] Marc R Dickey. A comparison of verbal instruction and nonverbal teacher-student modeling in instrumental ensembles. *J. Res. Music. Educ.*, 39(2):132–142, 1991.

[36] Robert A Duke. Measures of instructional effectiveness in music research. *Bull. Counc. Res. Music. Educ.*, pages 1–48, 1999.

[37] Shing-On Leung. A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point likert scales. *Journal of social service research*, 37(4):412–421, 2011.

[38] Kerry D Carlin. *Piano pedagogue perception of teaching effectiveness in the preadolescent elementary level applied piano lesson as a function of teacher behavior*. PhD thesis, Indiana University, 1997.

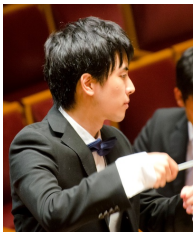
[39] Katie Zhukov. *Teaching styles and student behaviour in instrumental music lessons in*

Australian conservatoriums. PhD thesis, University of New South Wales, 2005.

[40] Sam Thompson and Aaron Williamon. Evaluating evaluation: Musical performance assessment as a research tool. *Music Perception*, 21(1):21–41, 2003.

[41] Brian C Wesolowski, Stefanie A Wind, and George Engelhard Jr. Examining rater precision in music performance assessment: An analysis of rating scale structure using the multifaceted rasch partial credit model. *Music Perception: An Interdisciplinary Journal*, 33(5):662–678, 2016.

Masaki Matsubara



Masaki Matsubara is an associate professor at Faculty of Library, Information and Media science, University of Tsukuba. His research interests include embodied cognitive science and art therapy, with a focus on promoting awareness of one's own ways of being. He has a Ph.D. in engineering from Keio University in 2013. He is a member of The Society for Art And Science.

Rina Kagawa

Rina Kagawa is a senior assistant professor in Faculty of Medicine, University of Tsukuba since 2018. Rina received a B.Sc. in Medicine from Keio University in 2012 and Ph.D. from the University of Tokyo in 2018. Rina's research interests include the interaction between human and big data.

Takeshi Hirano

Takeshi Hirano received his PhD in Medical Science from Osaka University, Japan, in 2013. His research area is motor control for human movement while playing a wind instrument. He is a member of the Japanese Society for Music Perception and Cognition.

Isao Tsuji

Isao Tsuji graduated from the Tokyo University of the Arts in 1982 and the Detmold Conservatory in 1987. He is a member of the Musicological Society of Japan.