What makes musicians infer teaching intentions?

Atsuko Tominaga[1,*], Günther Knoblich[1], & Natalie Sebanz[1]

[1] Department of Cognitive Science, Central European University, Quellenstraße 51, 1100 Vienna, Austria

[*] Corresponding author: Tominaga_Atsuko@phd.ceu.edu

Abstract

Perceiving pedagogical intentions is vital when learning skills from others. Our previous research demonstrated that expert pianists systematically modulated their sound so as to teach musical expressive techniques such as articulation and dynamics. For example, pianists played slower and exaggerated each technique when they had an intention to teach. Here we investigated whether the modulations that expert pianists produce when they intend to teach are also perceived by listeners as conveying pedagogical intentions. In the current study, musicians listened to piano recordings where a musical expressive technique of either articulation or dynamics was implemented. Half of the recordings was produced when pianists were instructed to play as if they were teaching the designated musical technique in a lesson (i.e., teaching recordings) whereas the other half was produced when pianists were instructed to play as if they were performing it in a concert (i.e., performing recordings). Participants were asked to judge whether each recording was produced for teaching purposes or not. We calculated the accuracy of participants' judgments to investigate whether they could distinguish teaching recordings from performing recordings. Also, recordings were quantified with regard to tempo, articulation and dynamic so that we could perform correlation and multiple regression analysis to investigate which features of piano performance made musicians infer teaching intentions. The findings in Experiment 1 with a simple musical scale demonstrated that the accuracy of musicians' judgments was significantly above chance. Also, it was found that slower tempo contributed to musicians' judgments as teaching regardless of the techniques. Moreover, performances with exaggeration for each technique (e.g., longer legato, shorter staccato, larger contrast between forte and piano) were more likely to be judged as teaching. Experiment 2 aimed to replicate the findings of Experiment 1 with a more naturalistic piece. As Experiment 1, the accuracy of musicians' judgments was significantly above chance. However, we only replicated that performances with exaggerated dynamics (in particular, larger contrast between forte and piano) were more likely to be judged as teaching. Taken together, loudness (dynamics) seems

to be reliably used to infer teaching intentions regardless of the complexity of a musical piece. Also, typical pedagogical behaviour such as slowing down may not be necessarily perceived as teaching performance when learning complex skills such as artistic expression.

Keywords: teaching, intention, skill transmission, musical expression

What makes musicians infer teaching intentions?

Introduction

Learning from others is one of the important elements of skill acquisition. Not only are we able to learn by observing and imitating others, but also we benefit greatly from interacting with others such as teachers and peers (Tomasello, Kruger, & Ratner, 1993). Adults are often being pedagogical to children in order to explain and transmit cultural conventions (Csibra & Gergely, 2009). Active teaching seems to play a crucial role not only to transit skills over generations but also to further develop sophisticated cultures, which cannot be achieved by one single individual or generation (Tennie, Call, & Tomasello, 2009). From a learner's perspective, it is important to identify informative teachers and infer teachers' expectations so that learners can acquire skills through interacting with teachers (Gweon, 2020; Veissière, Constant, Ramstead, Friston, & Kirmayer, 2020).

In pedagogical settings where teachers are supposed to convey useful information to learners, it has been found that teachers often modulate their behaviour for teaching purposes. For example, adults are likely to modulate their speech and action for infants so as to help infants acquire skills (e.g., Brand, Baldwin, & Ashburn, 2002; Saint-Georges et al., 2013). Also, some studies have revealed that even towards adult learners, people modulated their speech and action in the similar way as they did for infants (McEllin, Knoblich, & Sebanz, 2017; Uther, Knoll, & Burnham, 2007). Moreover, McEllin, Sebanz, and Knoblich (2018) demonstrated that people could identify informative intentions such as acting with others or teaching by relying on specific kinematics cues (e.g., velocity profiles of movements). These findings suggest that experts modified their speech and action to send teaching intentions and that novices could successfully perceive the intentions.

Tominaga, Knoblich & Sebanz (*submitted*) extended this line of research to expertise transmission where skills to be acquired are complex such as artistic expression. Based on the assumption that expert pianists are skilled at controlling their sound to communicate, we

investigated how expert pianists could modulate their performance when they were intending to teach musical expressive techniques of articulation (the smoothness of sound) and dynamics (the loudness of sound). The results demonstrated that expert pianists could successfully modulate their performance by playing slower or exaggerating relevant aspects of the performance (e.g., producing shorter staccato or making a larger contrast between forte and piano) to teach musical expressive techniques. Therefore, it seems that even in the domain of expertise, experts exhibit general pedagogical behaviours to communicate with their potential learners.

Here we investigated whether the modulations that expert pianists made when they intended to teach were also perceived by listeners as conveying pedagogical intentions. We assumed that listeners would be able to infer intentions only by listening to recordings because some research revealed that musical emotion is often communicated by sound alone and that listeners seem to be able to infer performers' intended emotions only by listening to recorded performances (e.g., Akkermans et al., 2019; Gabrielsson & Juslin, 1996). In the present study, musicians listened to piano recordings where a musical expressive technique of either articulation or dynamics was implemented. They were asked to judge whether each recording was produced for teaching purposes or not. Half of the recordings was produced when pianists were instructed to play as if they were teaching the designated musical technique in a lesson (i.e., teaching recordings) whereas the other half was produced when pianists were instructed to play as if they were performing it in a concert (i.e., performing recordings). First, we calculated the accuracy of participants' judgments to examine whether they could distinguish teaching recordings from performing recordings. Furthermore, recordings were quantified with regard to tempo, articulation and dynamic. We performed correlation and multiple regression analysis to examine relationships between performance features of the recordings and how likely people think the recordings were produced for teaching purposes. Various piano recordings were collected from our previous experiments (Tominaga et al., *submitted*). The recordings were performed by various pianists multiple

times, therefore the recordings contained many variations in terms of timing, articulation and dynamics.

If performers' intentions are successfully communicated with learners, musicians would be able to distinguish teaching recordings from performing recordings. Moreover, pedagogical behavioural features in our previous experiments such as slower demonstration and exaggerated performance should be identified and used to infer teaching intentions. We examined which features of piano performance were correlated with participants judgments as teaching. We conducted Experiment 1 with a simple musical scale and Experiment 2 with a more naturalistic piece of music to replicate the findings from Experiment 1.

Experiment 1

Methods

Participants

We recruited 21 participants who had at least six years of training in any musical instrument. They were able to read sheet music and knew two musical expressive techniques of articulation and dynamics. One participant was excluded due to an experimental error. Therefore, 20 participants (Female: 13) were included for data analysis and had 11.8 years of musical training on average ($SD = 5.62$). They were all right-handed with a mean age of 28.8 ($SD = 9.09$). All participants were recruited through an online participant platform (SONA system, https://www.sona-systems.com). The study (No. 2020/02) was approved by the Psychological Research Ethics Board (PREBO) CEU PU in Austria.

Apparatus

The experiment was programmed in Python 3.8.2 using the PsychoPy Python library (2020.2.4; https://www.psychopy.org/) on a Mac Book Pro with Mac OS X Catalina 10.15.6. Stimuli were played using the Mido Python library (1.2.9;

https://mido.readthedocs.io/en/latest/) on a Max/MSP patcher (8.1.7; https://cycling74.com/products/max). During the experiment, participants listened to the stimuli via headphones (Audio-Technica ATH-M50X).

Stimuli

We selected stimuli from our previous experiments (Tominaga et al., *submitted*). Stimuli were produced by actual pianists on a weighed Yamaha MIDI (Musical Instrument Digital Interface) digital piano and recorded as MIDI files. Multiple pianists played one piece of music with a musical expressive technique of either articulation (*Figure 1*, A) or dynamics (*Figure 1*, B). Articulation refers to the smoothness of sound, which is comprised of legato and staccato. Legato indicates smooth and connected sound whereas staccato indicates sharp and separate sound. Dynamics refers to the loudness of sound, which is comprised of forte and piano. Forte indicates loud sound while piano indicates soft sound. The piece was taken from "A Dozen a Day - Play with Ease in Many Keys" by Edna-Mae Burnam and modified for the experiment. The stimuli were performed around 80 quarter-beats per minute.

In Tominaga et al., (*submitted*), participants were asked to perform the piece with either articulation or dynamics in two different conditions. In the teaching condition, participants were instructed to perform the piece with the designated expressive technique as if they were teaching it to students (e.g., in a lesson). In the performing condition, participants were instructed to perform the piece with the designated expressive technique as if they were performing it to an audience (e.g., in a concert). In Tominaga et al. (*submitted*; Experiment 1), there were 453 valid performances (i.e., performances without any pitch errors) from the teaching condition and 436 valid performances from the performing condition. For the current experiment, 96 recordings were chosen from the valid performances. We randomly sampled 24 articulation recordings and 24 dynamics recordings from the teaching condition as well as 24 articulation recordings and 24 dynamics recordings from the performing condition. It is important to note that the recordings from the teaching

condition did not necessarily exhibit specific features of teaching performance that we found in our previous experiments (e.g., exaggeration) since we randomly sampled the performances.

Procedure

Upon arrival, participants read information sheet about the experiment and gave informed consent prior to participation. In the experiment, all instructions were displayed on a computer screen in front of the participants and an experimenter also explained the procedure. Participants were instructed that they were going to listen to piano recordings with one musical expressive technique of either articulation or dynamics, which were either produced as if a pianist were teaching the designated expressive technique to students (e.g., in a lesson) or as if a pianist were performing it to an audience (e.g., in a concert). In each trial, participants listened to one recording and were asked whether the recording was produced for teaching purposes or not. Participants responded by pressing either a yes (left arrow key) or no (right arrow key) button. While listening to each recording, sheet music, which corresponded to the recording, was shown on the screen in front of the participants (*Figure 2*).

There were two blocks and each block only included the recordings with one musical expressive technique of either articulation or dynamics. Each block consisted of four practice trials and 48 experimental trials. Each recording was evaluated only once in the experiment. All participants completed both blocks and the order of the blocks was counterbalanced across the participants. The order of the recordings was randomised within each block.

At the end of the experiment, participants filled in a questionnaire about their demographic information and experience in musical instruments.

Data analysis

Data processing and statistical analysis were performed in R version 4.0.5. Correlation analysis was performed with the standard *cor* function and regression models for multiple regression were fit with the standard *lm* function from the *stats* R package. Stimuli (MIDI files) were converted to numerical data in terms of time, pitch and velocity for the onset and offset of each note using the *tuneR* R package (https://cran.r-project.org/web/packages/tuneR/tuneR.pdf).

Accuracy. First, we examined whether participants could accurately distinguish the stimuli chosen from the teaching condition in Tominaga et al., (*submitted*; Experiment 1) as teaching and the stimuli chosen from the performing condition in Tominaga et al., (*submitted*; Experiment 1) as performing. We compared how accurate participants were against the chance level (50%). The correct responses were either pressing the yes button when listening to the teaching recordings or pressing the no button when listening to the performing recordings. The incorrect responses were either pressing the yes button when listening to the performing recordings or pressing the no button when listening to the teaching recordings.

Correlation and multiple regression. Stimuli were quantified with regard to tempo (interonset intervals; IOIs), articulation (key-overlap time; KOT), dynamics (key velocity; KV) and dynamics contrast (key velocity difference; KV-Diff) only for 16th notes. Interonset intervals are time intervals between onsets of adjacent notes. Larger IOIs indicate slower tempo while smaller IOIs indicate faster tempo. Key-overlap time is the time overlap between two adjacent notes, namely the difference between the offset time of the current note and the onset time of the ensuing note (e.g., Bresin & Battel, 2000). Positive KOT values indicate legato styles whereas negative KOT values indicate staccato styles. Key velocity is obtained from MIDI data to describe how fast a performer hit the key. Larger KV values indicate forte styles while smaller KV values indicate piano styles. Additionally, we also measured dynamics contrast where one subcomponent of the technique moves to the other (e.g., from forte to piano, from staccato to legato) to illustrate how much dynamics

contrast a performer made at transition points.

First, in order to investigate relationships between performance features (i.e., IOIs, KOT, KV, KV-Diff) and participants' judgments as teaching (i.e., what percentage of participants responded as "yes"), we performed correlation analysis. Second, we run multiple regression and entered all four performance features to examine which predictor significantly contributed to participants' judgments as teaching. Since articulation and dynamics were comprised of two opposite directional values (i.e., legato vs. staccato, forte vs. piano), we created four separate models, which considered only one subcomponent of either legato, staccato, forte or piano.

For articulation recordings, there were two models. The Legato model considered only legato parts of the stimuli. We entered the legato parts of KOT and KV and KV-Diff or transition points from legato to staccato. The Staccato model considered only staccato parts of the stimuli. We entered the staccato parts of KOT and KV and KV-Diff of transition points from staccato to legato. Similarly, there were two models for dynamics recordings. The Forte model considered only forte parts of the stimuli for KV and KOT, and transition points from forte to piano for KV-Diff. The Piano model considered only piano parts of the stimuli for KV and KOT, and transition points from piano to forte for KV-Diff. With regard to tempo (IOIs), there was only one value for each recording regardless of the subcomponents because tempo was consistent across the performance. Therefore, we entered the same tempo value for the Legato and Staccato models or the Forte and Piano models.

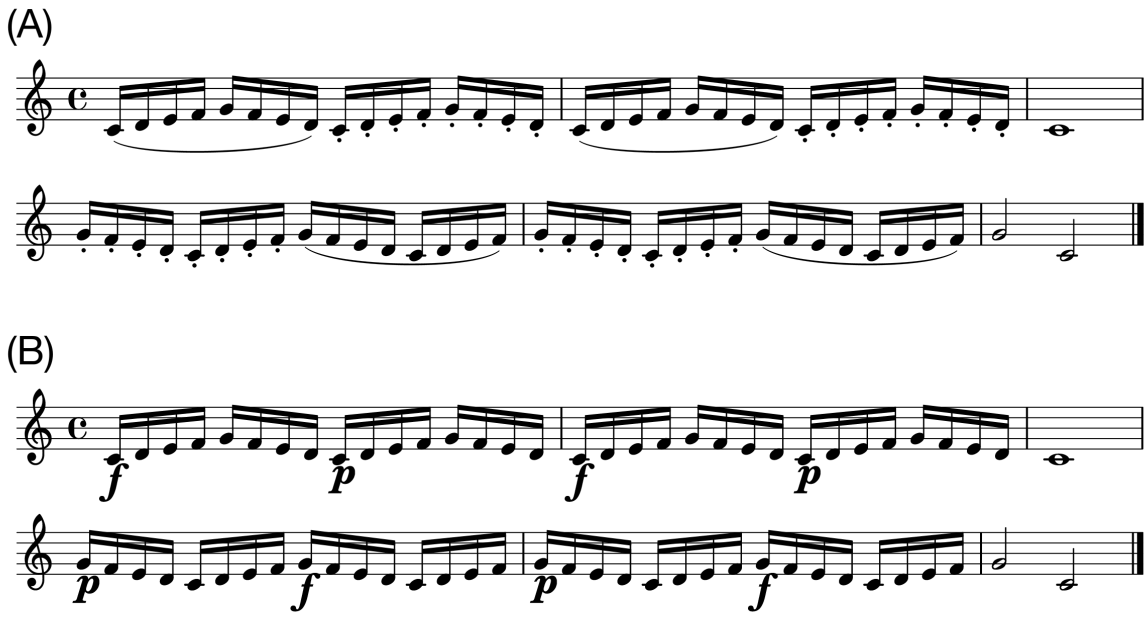

Figure 1. Stimuli. (A)Articulation. The curved line (slur) indicates legato and the dots indicate staccato. (B)Dynamics. The symbol 'f' denotes forte and the symbol 'p' denotes piano. Only the 16th notes were used for data analysis.

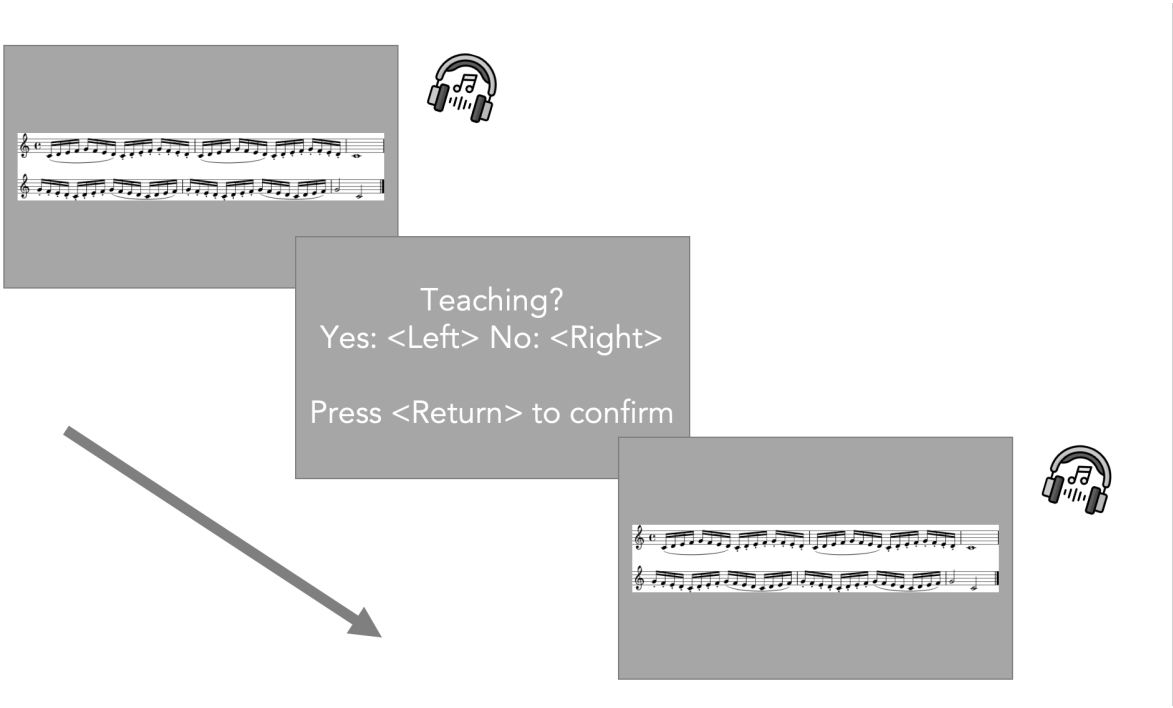

Figure 2. Procedure. Participants listened to a recording via headphones while corresponding sheet music was displayed on a monitor. They were required to respond by pressing a left-arrow (yes) or right-arrow (no) key for each judgment.

Results

All results were reported as significant at $p < 0.05$.

Accuracy

A one sample *t*-test was performed to compare the accuracy of participants' judgments against the chance level (50%). The mean percentage of correct answers [$M = 52.7$, $SD = 4.90$] was significantly higher than chance ($t(19) = 2.47$, $p = 0.02$, cohen's $d = 0.55$).

We also performed a one sample *t*-test for each technique separately. For articulation, the mean percentage of correct answers [$M = 50.2$, $SD = 6.55$] was not significantly different from chance ($t(19) = 0.14$, $p = 0.89$, cohen's $d = 0.03$). For dynamics, the mean percentage of correct answers [$M = 55.2$, $SD = 7.81$] was significantly higher than chance ($t(19) = 2.98$, $p = 0.01$, cohen's $d = 0.67$). A paired *t*-test revealed that there was a significant difference between the two techniques in terms of the accuracy ($t(19) = -2.12$, $p = 0.05$, cohen's $d = -0.69$), suggesting participants chose correct answers more for dynamics recordings than articulation recordings.

Correlation

A Pearson product-moment correlation coefficient was computed by default for correlation analysis and a Spearman's rank correlation coefficient was additionally computed if normality assumption was violated based on the Shapiro-Wilk normality test.

Tempo (IOIs). Performance tempi (IOIs) were significantly correlated with participants' judgments as teaching for both techniques (Articulation; $r(46) = .77$, $p < 0.001$, Dynamics; $r(46) = .42$, $p = 0.003$, *Figure 3*). Participants tended to identify slower performances as teaching.

However, the Shapiro-Wilk normality test revealed that the distribution of IOIs for dynamics recordings was not normally distributed ($p = 0.037$). Therefore, a Spearman's

rank correlation coefficient was additionally used to assess the relationship between performance tempi of dynamics recordings and participants' judgments. The result failed to reach significance ($r(46) = .27$, $p = 0.066$), suggesting that we need to be careful about the interpretation of the relationship between the performance tempi of dynamics recordings and participants' judgments as teaching.

Articulation (KOT). For articulation recordings, there was a significant relationship between KOT values and participants' judgments as teaching (*Figure 4*, left). Specifically, performances with shorter staccato ($r(46) = -.73$, $p < 0.001$) and longer legato ($r(46) = .40$, $p = 0.005$) were more likely to be judged as teaching.

For dynamics recordings, there was no significant relationship between KOT values and participants' judgments as teaching (Forte; $r(46) = -.04$, $p = 0.81$, Piano; $r(46) = -.19$, $p = 0.19$, *Figure 4*, right). The Shapiro-Wilk normality test revealed that the distribution of KOT for dynamics recordings was not normally distributed (Forte; $p < 0.001$, Piano; $p < 0.001$). Spearman's rank correlation coefficients also showed that there was no significant relationship between KOT values and participants' judgments as teaching (Forte; $r(46) = .15$, $p = 0.31$, Piano; $r(46) = -.12$, $p = 0.43$).

Dynamics (KV). For dynamics recordings, there was a significant relationship between KV values and participants' judgments as teaching (*Figure 5*, right). Specifically, performances with louder forte were more likely to be judged as teaching ($r(46) = .45$, $p = 0.001$). However, there was no significant relationship between KV values for piano and participants' judgments as teaching ($r(46) = -.22$, $p = 0.13$).

For articulation recordings, there was no significant relationship between KV values and participants' judgments as teaching (Legato; $r(46) = .10$, $p = 0.52$, Staccato; $r(46) = -.02$, $p = 0.87$, *Figure 5*, left). The Shapiro-Wilk normality test revealed that the distribution of KOT for articulation recordings (legato parts) was not normally distributed ($p = 0.002$). A Spearman's rank correlation coefficients also showed that there was no significant

relationship between KOT values for legato and participants' judgments as teaching ($r(46) = .03$, $p = 0.84$).

Dynamics contrast (KV-Diff). For dynamics recordings, there was a significant relationship between KV difference between forte and piano and participants' judgments as teaching (*Figure 6*, right). Specifically, performances with larger contrasts between forte and piano were more likely to be judged as teaching (From Forte to Piano; $r(46) = -.50$, $p < 0.001$, From Piano to Forte; $r(46) = .62$, $p < 0.001$).

For articulation recordings, there was no significant relationship between KV difference between legato and staccato and participants' judgments as teaching (From Legato to Staccato; $r(46) = -.26$, $p = 0.071$, From Staccato to Legato; $r(46) = -.19$, $p = 0.21$, *Figure 6*, left). The Shapiro-Wilk normality test revealed that the distribution of KV difference for articulation recordings (transition points from legato to staccato) was not normally distributed ($p = 0.007$). A Spearman's rank correlation coefficients also showed that there was no significant relationship between KOT values for the transition points from legato to staccato and participants' judgments as teaching ($r(46) = -.23$, $p = 0.12$).

Multiple regression

In order to further explore which feature of performance contributed the most to participants' judgments as teaching, multiple regression analyses were conducted. Statistical model assumptions were tested using the *performance* R package (Lüdecke, Ben-Shachar, Patil, Waggoner, & Makowski, 2021) and most of the assumptions (see *Supplementary Material* for details) were met. Since articulation and dynamics consisted of two opposite subcomponents (i.e., legato vs. staccato, forte vs. piano) and therefore cannot be summed up to represent each technique as one value, we reported four separate regression models for each subcomponent (see details in *Data analysis*).

Legato. A multiple regression analysis was conducted to predict participants' judgments as teaching based on performance features of tempo (IOIs), articulation (KOT for

legato parts), dynamics (KV for legato parts) and dynamics contrast (KV-Diff from legato to staccato). The result of the regression indicated that the model explained 64.6 % of the variance ($F(4, 43) = 22.5$, $p < 0.001$). It was found that tempo (IOIs; $\beta = 0.78$, $p < 0.001$) and articulation for the legato parts (KOT; $\beta = 0.26$, $p = 0.004$) were significant predictors of participants' judgments as teaching.

Staccato. A multiple regression analysis was conducted to predict participants' judgments as teaching based on performance features of tempo (IOIs), articulation (KOT for staccato parts), dynamics (KV for staccato parts) and dynamics contrast (KV-Diff from staccato to legato). The result of the regression indicated that the model explained 64.0 % of the variance ($F(4, 43) = 21.9$, $p < 0.001$). It was found that tempo (IOIs; $\beta = 0.52$, $p = 0.002$) and articulation for the staccato parts (KOT; $\beta = -0.28$, $p = 0.020$) were significant predictors of participants' judgments as teaching.

Forte. A multiple regression analysis was conducted to predict participants' judgments as teaching based on performance features of tempo (IOIs), articulation (KOT for forte parts), dynamics (KV for forte parts) and dynamics contrast (KV-Diff from forte to piano). The result of the regression indicated that the model explained 35.9 % of the variance ($F(4, 43) = 7.58$, $p < 0.001$). It was found that tempo (IOIs; $\beta = 0.35$, $p = 0.007$) and dynamics for the forte parts (KV; $\beta = 0.73$, $p = 0.048$) were significant predictors of participants' judgments as teaching.

Piano. A multiple regression analysis was conducted to predict participants' judgments as teaching based on performance features of tempo (IOIs), articulation (KOT for piano parts), dynamics (KV for piano parts) and dynamics contrast (KV-Diff from piano to forte). The result of the regression indicated that the model explained 51.2 % of the variance ($F(4, 43) = 13.3$, $p < 0.001$). It was found that tempo (IOIs; $\beta = 0.38$, $p < 0.001$) and dynamics contrast from piano to forte (KV-Diff; $\beta = 0.99$, $p < 0.001$) were significant predictors of participants' judgments as teaching.

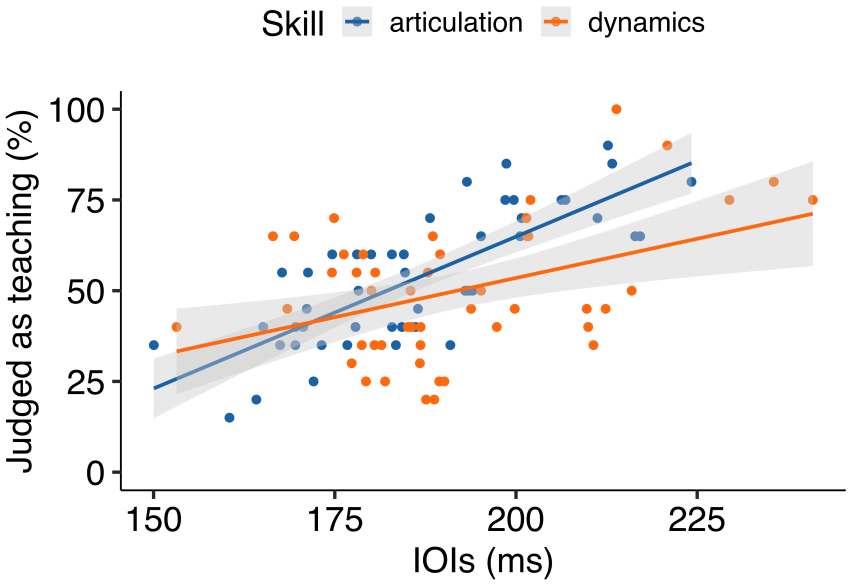

Figure 3. Experiment 1: Scatter plot showing the correlation between tempo feature (IOIs) and average participants' judgments as teaching for each recording. Therefore, each dot represents each stimulus.

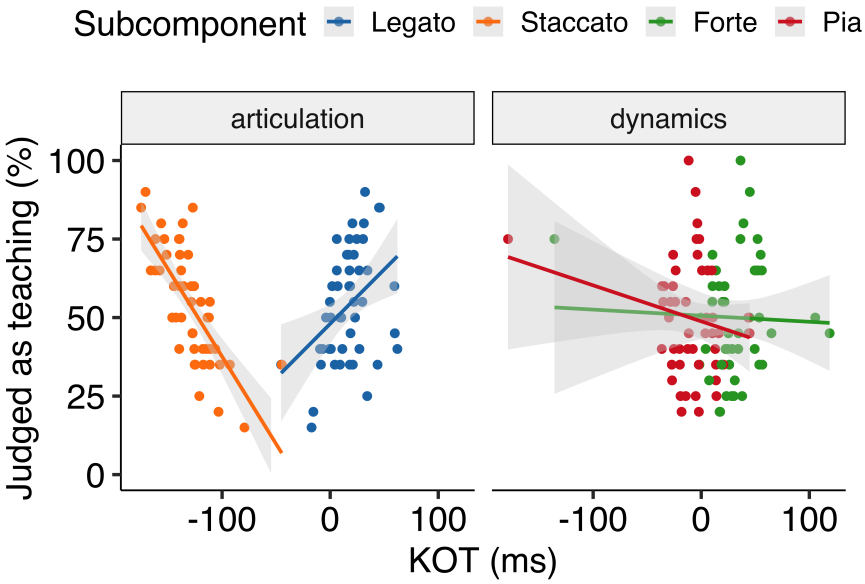

Figure 4. Experiment 1: Scatter plot showing the correlation between articulation feature (KOT) and average participants' judgments as teaching for each recording. Therefore, each dot represents each stimulus.

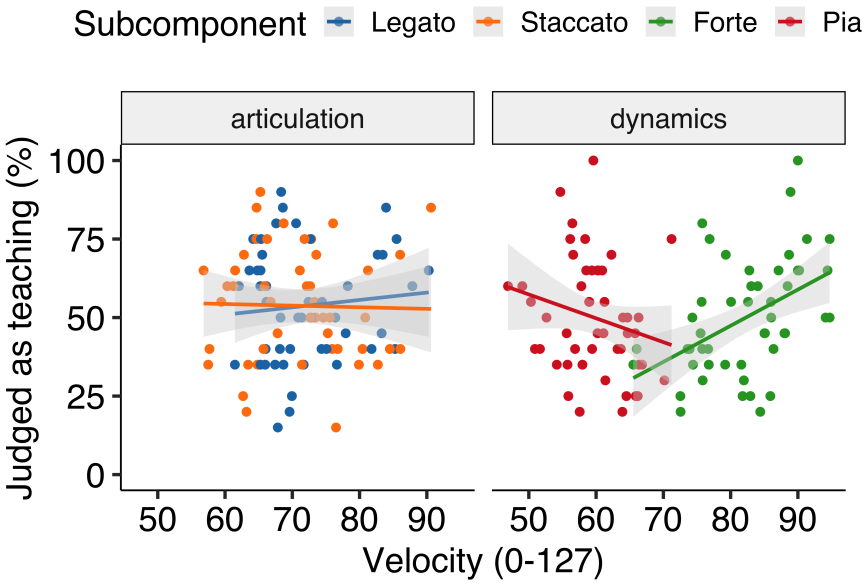

Figure 5. Experiment 1: Scatter plot showing the correlation between dynamics feature (KV) and average participants' judgments as teaching for each recording. Therefore, each dot represents each stimulus.

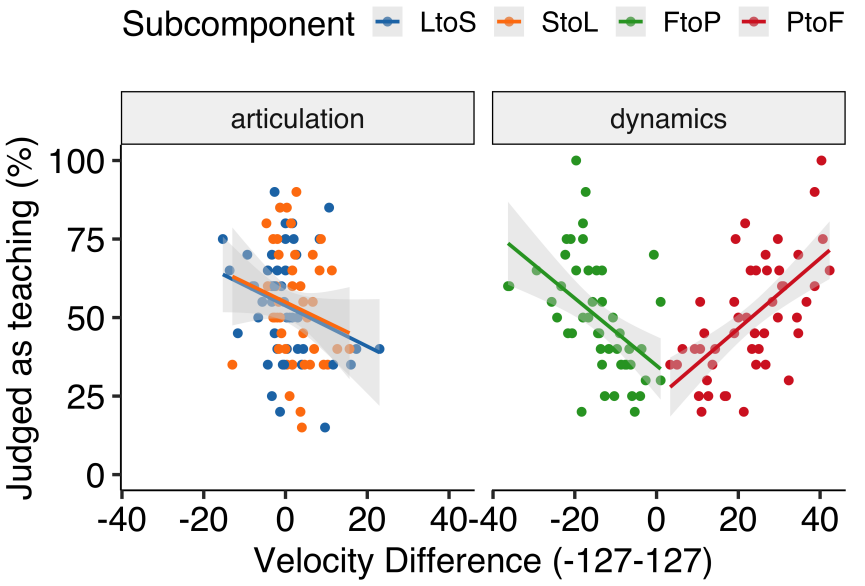

Figure 6. Experiment 1: Scatter plot showing the correlation between dynamics contrast feature (KV-Diff) and average participants' judgments as teaching for each recording. Therefore, each dot represents each stimulus.

Discussion

Experiment 1 investigated whether musicians could distinguish teaching recordings from performing recordings accurately and which features of piano performance made them infer teaching intentions. The results demonstrated that musicians could choose correct answers above chance. Also, it was found that the accuracy of participants' judgments was better for dynamics recordings than for articulation recordings. When it comes to how performance features contributed to participants' judments, performances with slower tempo were more likely to be judged as teaching by musicians regardless of which expressive technique was implemented in the piece. For articulation recordings, performances with longer legato and shorter staccato were tended to be judged as teaching. For dynamics recordings, performances with louder forte were more likely to be judged as teaching whereas there was no relationship between softer sound (i.e., piano) and participants' judgments as teaching. Importantly, performances with larger contrast between forte and piano for both directions (i.e., from forte to piano, from piano to forte) were more likely to be judged as teaching. This result may suggest that dynamics contrast might be reliably used to infer teaching intentions, rather than absolute dynamics values themselves. Moreover, multiple regression analyses implied that tempo feature was the strongest predictor of participants' judgments as teaching in general whereas there were specific predictors depending on which expressive technique was implemented in the piece. These performance features were overall consistent with what expert pianists did in our previous experiments for teaching purposes. Therefore, our findings suggest that musicians may rely on generic pedagogical behaviours (e.g., slower demonstration, exaggeration) to infer teaching intentions of expert pianists, only by listening to recorded performances.

Experiment 2

The aim of Experiment 2 was to replicate the findings in Experiment 1 with a more naturalistic piece of music. Given the findings in Experiment 1, we predicted that slower performance would be likely to be judged as teaching regardless of which expressive technique (articulation or dynamics) was implemented. Also performances with exaggerated articulation and dynamics (in particular, longer legato and shorter staccato, larger contrast between forte and piano) would be likely to be judged as teaching.

Methods

Participants

We recruited 21 participants who had at least six years of training in any musical instrument or singing. They were able to read sheet music and know two musical expressive techniques of articulation and dynamics. One participant was excluded because s/he did not understand the instructions. Therefore, 20 participants (Female: 10) were included for data analysis and had 12.65 years of training on average in any musical instrument or singing (SD = 5.40). Most people were right-handed (Left; 1) with a mean age of 33.55 (SD = 12.80). As Experiment 1, all participants were recruited through the SONA system and the study (No. 2020/02) was approved by the PREBO CEU PU in Austria.

Apparatus and procedure

The apparatus and procedure were identical to Experiment 1 except that each block consisted of four practice trials and 36 experimental trials. The number of trials was reduced due to the time constraint of the experiment.

Stimuli

As Experiment 1, we selected stimuli from our previous experiments (Tominaga et al., *submitted*; Experiment 2). The excerpt was taken from "Sonatina Op.36 (No.3) in C major"

by Muzio Clementi and modified for the experiment. The excerpt was performed with either articulation (*Figure 7*, A) or dynamics (*Figure 7*, B). The stimuli were performed around 100 - 120 quarter-beats per minute.

For the current experiment, 72 performances were chosen from the valid performances in Tominaga et al. (*submitted*; Experiment 2). There were 248 valid performances in the teaching condition and 256 valid performances in the performing condition. We randomly sampled 18 articulation performances and 18 dynamics performances from the teaching condition as well as 18 articulation performances and 18 dynamics performances from the performing condition. Again, it is important to note that each performance from the teaching condition did not necessarily exhibit specific features of teaching that we found in the previous experiments (e.g., exaggeration) since we randomly sampled the performances.

Data analysis

The data analysis was almost identical to Experiment 1. Only the 8th notes with expressive notations were included for data analysis. As a result, only one 8th note in the 4th measure without any expression was not included.

Results

All results were reported as significant at $p < 0.05$.

Accuracy

A one sample t-test was performed to compare the accuracy of participants' judgments against the chance level (50%). The mean percentage of correct answers [$M = 52.8$, $SD = 4.78$] was significantly higher than chance ($t(19) = 2.66$, $p = 0.02$, cohen's $d = 0.60$).

We also performed a one sample t-test for each technique separately. For articulation, the mean percentage of correct answers [$M = 52.4$, $SD = 6.94$] was not significantly different from chance ($t(19) = 1.52$, $p = 0.14$, cohen's $d = 0.34$). Also for dynamics, the mean

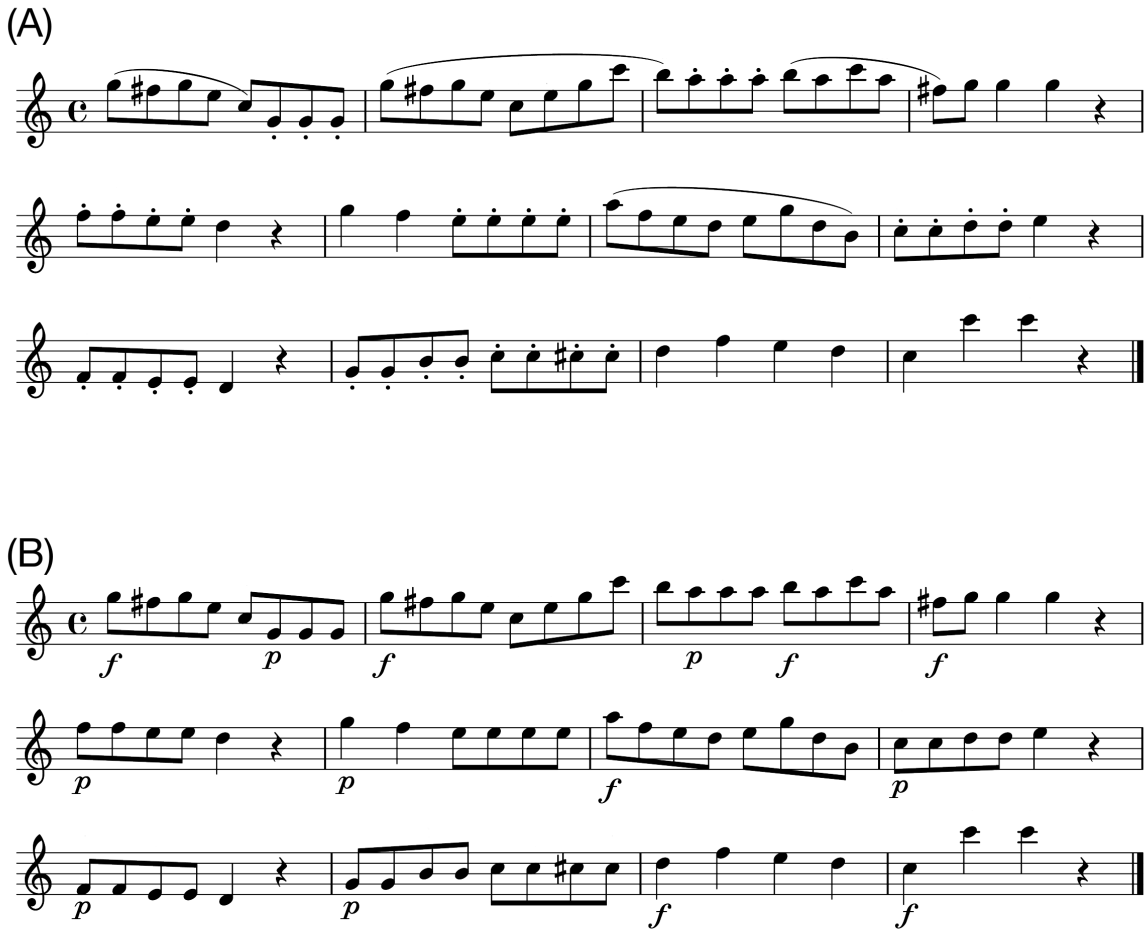

Figure 7. Stimuli. (A)Articulation. The curved line (slur) indicates legato and the dots indicate staccato. (B)Dynamics. The symbol 'f' denotes forte and the symbol 'p' denotes piano. Only the 8th notes with expressive notations were used for data analysis.

percentage of correct answers [M = 53.3, SD = 8.53] was not significantly different from chance ($t(19)$ = 1.75, p = 0.10, cohen's d = 0.39). A paired t-test revealed that there was no significant difference between the two techniques in terms of the accuracy ($t(19)$ = -0.35, p = 0.73, cohen's d = -0.13).

Correlation

As Experiment 1, a Pearson product-moment correlation coefficient was computed by default for correlation analysis and a Spearman's rank correlation coefficient was additionally computed if normality assumption was violated based on the Shapiro-Wilk normality test.

Tempo (IOIs). Unlike Experiment 1, there was a significant relationship between performance tempi (IOIs) and participants' judgments as teaching only for dynamics recordings (Articulation; *r(34)* = .25, p = 0.15, Dynamics; *r(34)* = .39, p = 0.02, *Figure 8*). Participants tended to identify slower performances as teaching for dynamics recordings.

However, the Shapiro-Wilk normality test revealed that the distribution of IOIs for dynamics recordings was not normally distributed (p = 0.080). A Spearman's rank correlation coefficient also showed that there was no significant relationship between tempi for dynamics performances and participants' judgments as teaching (*r(34)* = .29, p = 0.091), suggesting that we need to be careful about the interpretation of the relationship between the performance tempi of dynamics recordings and participants' judgments as teaching.

Articulation (KOT). Unlike Experiment 1, for articulation recordings, there was no significant relationship between KOT values and participants' judgments as teaching (Legato; *r(34)* = -.03, p = 0.39, Staccato; *r(34)* = -.15, p = 0.39, *Figure 9*, left).

For dynamics recordings, there was no significant relationship between KOT values for forte and participants' judgments as teaching (*r(34)* = -.11, p = 0.52). However, there was a significant relationship between KOT values for piano and participants' judgments as teaching (*r(34)* = -.35, p = 0.034), suggesting that performances with staccato-style piano

were more likely to be considered as teaching performance (*Figure 9*, right).

Dynamics (KV). As Experiment 1, for dynamics recordings, there was a significant relationship between KV values and participants' judgments as teaching (*Figure 10*, right). Specifically, performances with louder forte ($r(34) = .45$, $p = 0.007$) and softer piano ($r(34) = -.45$, $p = 0.006$) were more likely to be judged as teaching. The Shapiro-Wilk normality test revealed that the distribution of KV values for dynamics recordings was not normally distributed (Forte; $p = 0.048$, Piano' $p = 0.004$). Spearman's rank correlation coefficient also confirmed that performances with louder forte ($r(34) = .43$, $p = 0.010$) and softer piano ($r(34) = -.45$, $p = 0.013$) were more likely to be judged as teaching.

For articulation recordings, there was no significant relationship between KV values and participants' judgments as teaching (Legato; $r(34) = .08$, $p = 0.63$, Staccato; $r(34) = .21$, $p = 0.22$, *Figure 10*, left).

Dynamics contrast (KV-Diff). As Experiment 1, for dynamics recordings, there was a significant relationship between KV difference between forte and piano and participants' judgments as teaching (*Figure 11*, right). Specifically, performances with larger contrasts between forte and piano were more likely to be judged as teaching (From Forte to Piano; $r(34) = -.75$, $p < 0.001$, From Piano to Forte; $r(34) = .59$, $p < 0.001$). The Shapiro-Wilk normality test revealed that the distribution of KV-Diff values for dynamics recordings was not normally distributed (Forte to Piano; $p = 0.85$, Piano to Forte; $p = 0.008$). Spearman's rank correlation coefficient also confirmed that performances with larger contrasts between forte and piano were more likely to be judged as teaching (Forte to Piano; $r(34) = -.73$, $p < 0.001$, Piano to Forte; $r(34) = .57$, $p < 0.001$).

For articulation recordings, there was no significant relationship between KV difference between transition points from legato to staccato and participants' judgments as teaching ($r(34) = .23$, $p = 0.176$). However, there was a significant relationship between the transition points from staccato to legato and participants' judgments as teaching ($r(34) = .36$, $p =$

0.03), suggesting that performances with larger contrast from staccato to legato were more likely to be considered as teaching performance (*Figure 11*, left).

Multiple regression

As Experiment 1, we performed multiple regression analyses to further explore which feature of performance contributed the most to participants' judgments as teaching. Statistical model assumptions were tested and most of the assumptions (see *Supplementary Material* for details) were met. Again, since articulation and dynamics consisted of two opposite subcomponents (i.e., legato vs. staccato, forte vs. piano) and therefore cannot be summed up to represent each technique as one value, we reported four separate regression models for each subcomponent (see details in *Data analysis* in Experiment 1).

Legato. A multiple regression analysis was conducted to predict participants' judgments as teaching based on performance features of tempo (IOIs), articulation (KOT for legato parts), dynamics (KV for legato parts) and dynamics contrast (KV-Diff from legato to staccato). The result of the regression indicated that the model explained 19.8 % of the variance ($F(4, 31) = 3.16$, $p = 0.028$). It was found that tempo (IOIs; $\beta = 0.58$, $p = 0.011$), dynamics (KV; $\beta = 1.10$, $p = 0.018$) and dynamics contrast (KV-Diff; $\beta = 1.90$, $p = 0.006$) for the legato parts were articulation for the legato parts were significant predictors of participants' judgments as teaching. However, articulation (KOT; $\beta = -0.08$, $p = 0.47$) for the legato parts was not a significant predictor as opposed to Experiment 1.

Staccato. A multiple regression analysis was conducted to predict participants' judgments as teaching based on performance features of tempo (IOIs), articulation (KOT for staccato parts), dynamics (KV for staccato parts) and dynamics contrast (KV-Diff from staccato to legato). The overall model was not statistically significant ($R^2 = 12.3$, $F(4, 31) = 2.22$, $p = 0.09$).

Forte. A multiple regression analysis was conducted to predict participants' judgments as teaching based on performance features of tempo (IOIs), articulation (KOT for

forte parts), dynamics (KV for forte parts) and dynamics contrast (KV-Diff from forte to piano). The result of the regression indicated that the model explained 60.7 % of the variance ($F(4, 31) = 14.5$, $p < 0.001$). It was found that dynamics contrast from forte to piano (KV-Diff; β = -1.64, $p < 0.001$) were significant predictors of participants' judgments as teaching.

Piano. A multiple regression analysis was conducted to predict participants' judgments as teaching based on performance features of tempo (IOIs), articulation (KOT for piano parts), dynamics (KV for piano parts) and dynamics contrast (KV-Diff from piano to forte). The result of the regression indicated that the model explained 49.5 % of the variance ($F(4, 31) = 9.57$, $p < 0.001$). It was found that tempo (IOIs; $\beta = 0.55$, $p = 0.022$) and dynamics contrast from piano to forte (KV-Diff; $\beta = 1.09$, $p < 0.001$) were significant predictors of participants' judgments as teaching.

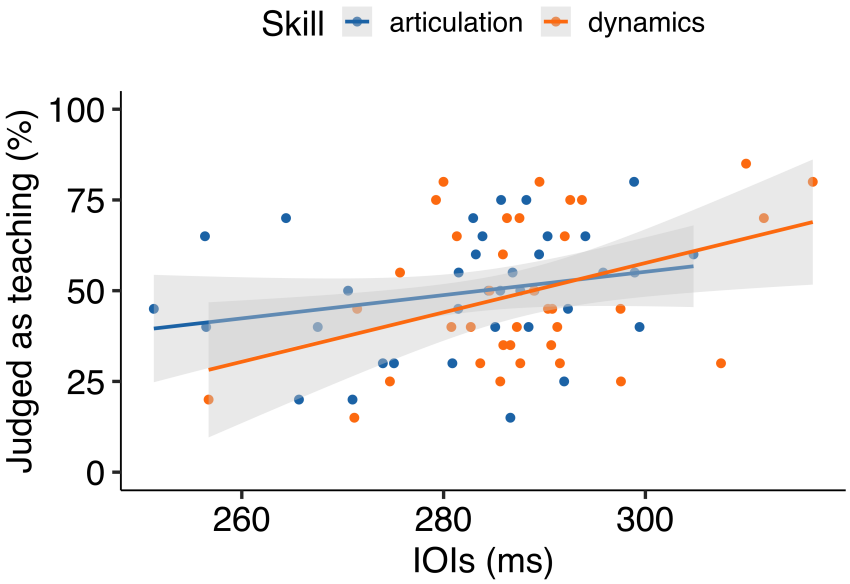

Figure 8. Experiment 2: Scatter plot showing the correlation between tempo featurea (IOIs) and average participants' judgments as teaching for each recording. Therefore, each dot represents each stimulus.

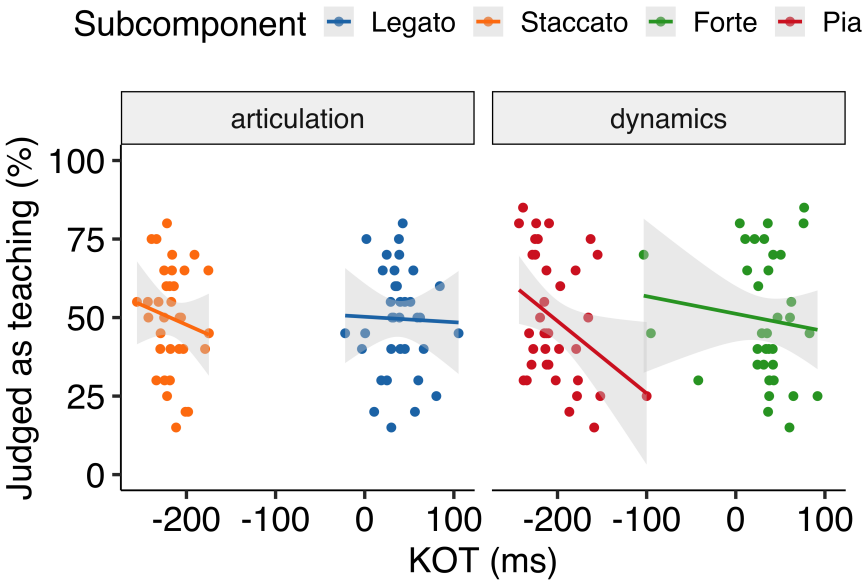

Figure 9. Experiment 2: Scatter plot showing the correlation between articulation features (KOT) and average participants' judgments as teaching for each recording. Therefore, each dot represents each stimulus.

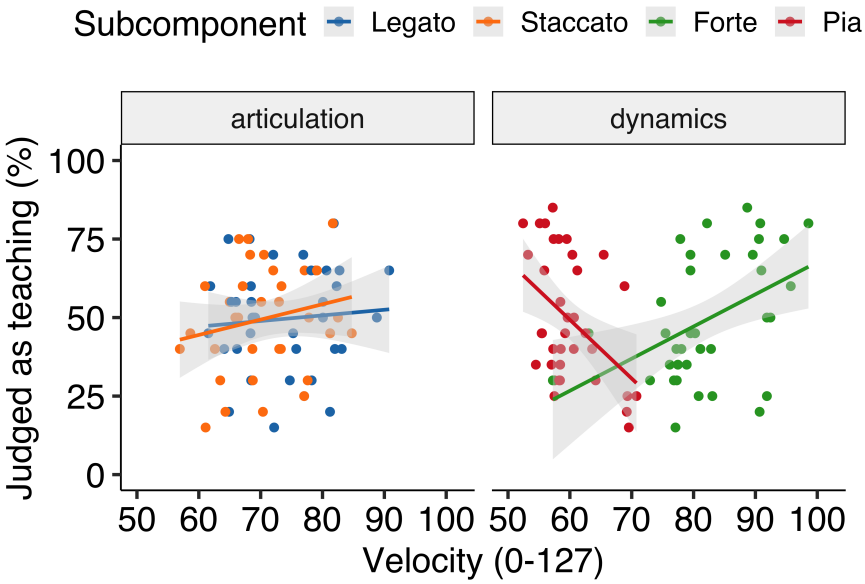

Figure 10. Experiment 2: Scatter plot showing the correlation between dynamics features (KV) and average participants' judgments as teaching for each recording. Therefore, each dot represents each stimulus.

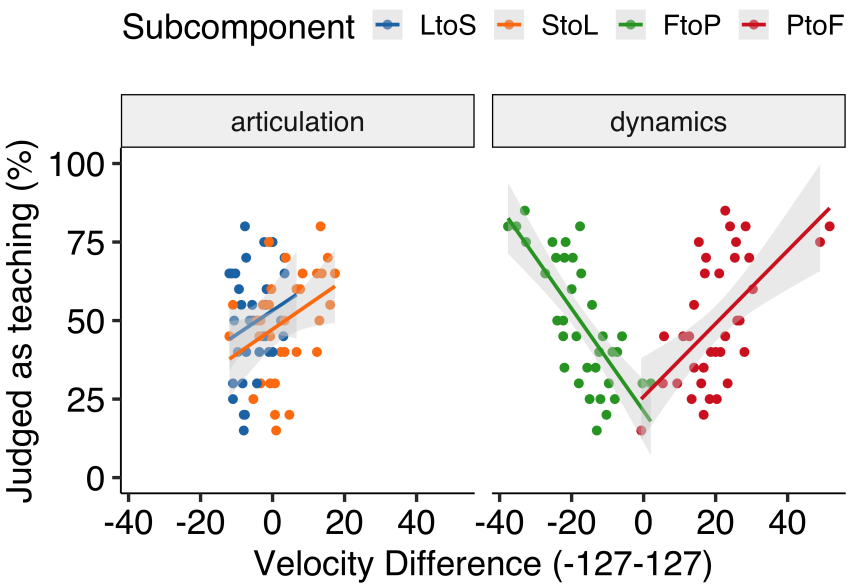

Figure 11. Experiment 2: Scatter plot showing the correlation between dynamics contrast features (KV-Diff) and average participants' judgment as teaching for each stimulus.

Discussion

The aim of Experiment 2 was to replicate the results of Experiment 1 with a more naturalistic piece of music. However, unlike Experiment 1, tempo (IOIs) does not seem to contribute to participants' judgments as teaching. This could be that in a musically complex piece, tempo might have been considered to be one of the interpretations of music and was not used as a reliable cue to infer teaching intentions. Also, articulation (KOT) does not seem to contribute to participants' judgments as teaching. As Experiment 1, dynamics (KV) seems to be used as a reliable cue to detect the characteristic of teaching performance. We could successfully replicate that performances with larger contrast between forte and piano seem to be considered as teaching. Moreover, performances with both exaggerated forte and piano were considered to be for teaching. These consistent results suggest that loudness (dynamics) might be used as a reliable cue to infer teaching intentions regardless of the complexity of a musical piece.

Multiple regression analyses also implied that tempo feature was not a strong predictor of participants' judgments as teaching when listening to a musically complex piece. However, dynamics contrast seems to be a strong predictor of participants' judgments as teaching, especially for dynamics recordings. This finding also suggests that loudness might be a strong cue to infer pedagogical intentions even in a naturalistic piece of music.

General discussion

The present study investigated whether and how musician infer pedagogical intentions by listening to piano recordings. In both Experiment 1 and 2, participants were able to choose correct answers accurately more than chance (50%). However, the results showed that the accuracy was relatively low (52.7 % in Experiment 1; 52.8 % in Experiment 2). This might be because the stimuli were randomly sampled from the teaching and performing condition in Tominaga et al. (*submitted*) and all the selected recordings did not exhibit the characteristics of teaching performance (e.g., slow performance, exaggeration). The reason why we randomly selected the stimuli from each condition (i.e., teaching or performing) was because all the characteristics of teaching performance (e.g., IOIs, KOT, KV) were intercorrelated and it was difficult to choose stimuli based on one single criterion. Future research should examine if musicians answer more accurately with more experimentally controlled stimuli.

In order to examine which features of piano performance make musicians infer pedagogical intentions, we performed correlation and multiple regression analysis. Across the two experiments, it was found that loudness, particularly larger contrast between forte and piano, strongly contributed to participants' judgments when listening to dynamics recordings. On the other hand, unlike what we predicted, slower performance was only considered to be for teaching in Experiment 1 where the stimuli consisted of a simple musical scale. Also, we could not find that performances with exaggerated articulation (e.g., longer legato and shorter staccato) were more likely to be judged as teaching in Experiment 2. This might

suggest that some characteristics of performance are not necessarily used or reliable to infer performers' intentions when listening to musically complex pieces. It can be speculated that certain modulations are too subtle so that participants could not perceive them. Another possibility is that modulations can also be considered as one of the expressions or interpretations.

In the current study, we exclusively recruited musicians to explore our research questions. The reason why we recruited musicians only was because the concepts of articulation and dynamics seem to be difficult for those who don't play an instrument to understand in the current settings. It would be important to investigate how novices perceive and infer pedagogical intentions differently from musicians, who already have some experience in playing music.

References

Akkermans, J., Schapiro, R., Müllensiefen, D., Jakubowski, K., Shanahan, D., Baker, D., ... Frieler, K. (2019). Decoding emotions in expressive music performances: A multi-lab replication and extension study. *Cognition and Emotion*, *33*(6), 1099–1118. https://doi.org/10.1080/02699931.2018.1541312

Brand, R. J., Baldwin, D. A., & Ashburn, L. A. (2002). Evidence for "motionese": Modifications in mothers' infant-directed action. *Developmental Science*, *5*(1), 72–83. https://doi.org/10.1111/1467-7687.00211

Bresin, R., & Battel, G. U. (2000). Articulation Strategies in Expressive Piano Performance Analysis of Legato, Staccato, and Repeated Notes in Performances of the Andante Movement of Mozart's Sonata in G Major (K 545). *Journal of New Music Research*, *29*(3), 211–224. https://doi.org/10.1076/jnmr.29.3.211.3092

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences*, *13*(4), 148–153. https://doi.org/10.1016/j.tics.2009.01.005

Gabrielsson, A., & Juslin, P. N. (1996). Emotional Expression in Music Performance: Between the Performer's Intention and the Listener's Experience. *Psychology of Music*, *24*(1), 68–91. https://doi.org/10.1177/0305735696241007

Gweon, H. (2020). The role of communication in acquisition, curation, and transmission of culture. *Behavioral and Brain Sciences*, *43*. https://doi.org/10.1017/S0140525X19002863

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). Performance: An R Package for Assessment, Comparison and Testing of Statistical Models. *Journal of Open Source Software*, *6*(60), 3139. https://doi.org/10.21105/joss.03139

McEllin, L., Knoblich, G., & Sebanz, N. (2017). Distinct kinematic markers of demonstration and joint action coordination? Evidence from virtual xylophone playing. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(6), 885. https://doi.org/10.1037/xhp0000505

McEllin, L., Sebanz, N., & Knoblich, G. (2018). Identifying others' informative intentions from movement kinematics. *Cognition*, *180*, 246–258. https://doi.org/10.1016/j.cognition.2018.08.001

Saint-Georges, C., Chetouani, M., Cassel, R., Apicella, F., Mahdhaoui, A., Muratori, F., ... Cohen, D. (2013). Motherese in Interaction: At the Cross-Road of Emotion and Cognition? (A Systematic Review). *PLOS ONE*, *8*(10), e78103. https://doi.org/10.1371/journal.pone.0078103

Tennie, C., Call, J., & Tomasello, M. (2009). Ratcheting up the ratchet: On the evolution of cumulative culture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1528), 2405–2415. https://doi.org/10.1098/rstb.2009.0052

Tomasello, M., Kruger, A. C., & Ratner, H. H. (1993). Cultural learning. *Behavioral and Brain Sciences*, *16*(3), 495–511. https://doi.org/10.1017/S0140525X0003123X

Uther, M., Knoll, M. A., & Burnham, D. (2007). Do you speak E-NG-L-I-SH? A comparison of foreigner- and infant-directed speech. *Speech Communication*, *49*(1), 2–7. https://doi.org/10.1016/j.specom.2006.10.003

Veissière, S. P. L., Constant, A., Ramstead, M. J. D., Friston, K. J., & Kirmayer, L. J. (2020). Thinking through other minds: A variational approach to cognition and culture. *Behavioral and Brain Sciences*, *43*. https://doi.org/10.1017/S0140525X19001213