The effect of contextual interference and feedback on learning composers' musical styles

[1]The Department of Physiology and Pharmacology, Schulich School of Medicine & Dentistry, The University of Western Ontario, London, ON, Canada

[2]Brain and Mind Institute, The University of Western Ontario, London, ON, Canada

[3]Department of Psychology, The University of Western Ontario, London, ON, Canada`

Table of Contents

Abstract

Introduction: Musicians most commonly learn to distinguish musical styles through repeated listening in long blocks. Although blocked practice often seems like the most intuitive way to learn, alternating between tasks using an interleaved practice schedule demands more effort and has been shown to improve retention, a phenomenon known as the “contextual interference effect”. However, there are few studies of contextual interference on auditory category learning, nor has it been tested in conjunction with other learning strategies, such as feedback. The current study hypothesized that the combination of contextual interference and feedback would additively benefit learning of composers’ musical styles.

Methods: In a two-day online study, 134 participants of varying musical backgrounds listened to music excerpts from three composers in a blocked schedule (one composer presented consecutively) and excerpts from another three composers in an interleaved schedule (various composers alternated). Participants then classified novel excerpts from the same six composers with or without corrective feedback. On the second day, participants were tested on another set of novel excerpts and completed a metacognitive judgment questionnaire assessing in which condition they felt they learned best. Effects of practice schedule and feedback on test performance were analyzed using a two-way mixed analysis of variance (ANOVA).

Results: In the no feedback group, performance was similar across blocked and interleaved conditions. In the feedback group, interleaving produced significantly greater accuracy than blocking (interaction: $p<0.05$). Yet, the majority of participants misjudged blocking to be more effective than interleaving.

Conclusions: Combining an interleaved practice schedule and feedback were both required to improve test performance. These findings suggest that supplementing contextual interference with feedback enhances category learning in the auditory domain, which has practical implications for improving music education.

Keywords: contextual interference, interleaving, music learning, inductive category learning, metacognition

Introduction

In music, the most widely adopted form of practice is repetition in which musicians are encouraged to play through music passages multiple times in a row (Austin and Berg, 2006; Maynard, 2006). In psychology, this technique is called blocked practice, where one task is practiced repetitively before moving onto the second task in an AAABBBCCC fashion (Williams, 2006). Blocking is an intuitive strategy, as repeating an individual task increases feelings of fluency and increases performance during a practice session (Simon, 2007). Extensive research, however, has demonstrated that blocked practice is not optimal for improving performance after practice (Magill and Hall, 1990; Carter and Grahn, 2016).

Compared to focusing exclusively on one task at a time, alternating between different tasks (e.g. ABCACBCAB) can lead to better memory (Magill and Hall, 1990). In baseball batting, for example, hitting 15 fastballs in a row, 15 curve-balls and then 15 change-up pitches is a blocking strategy that may facilitate performance and feelings of fluency during practice. However, alternating the 30 different pitches in a random order leads to superior performance after practice and during an actual game, when the real test of learning occurs (Hall et al., 1994). This alternating style of practice is called interleaved practice.

Interleaved practice creates interference that reduces performance while practicing. Although making practice more challenging may seem counterintuitive, interleaving places greater demands on the working memory by limiting habituation and encouraging more effortful processing (Magill and Hall, 1990). Ultimately, this leads to improved retention and long-term performance, a phenomenon known as the contextual

interference effect (Magill and Hall, 1990). Overall, as more desirable difficulties and interferences are introduced, contextual interference increases and causes poorer performance in the short-term during early learning but improves long-term retention (Bjork, 1994). This phenomenon was first observed in word-pair learning studies by Battig (1966) and later demonstrated in motor skill learning by Shea and Morgan (1979). Since then, the contextual interference effect has been extensively studied in experiments examining practice schedules for motor skills and sports psychology (Magill and Hall, 1990). For instance, interleaved practice improves training in baseball, badminton, and basketball, among other sports (Goode and Magill, 1986; Hall et al., 1994; Feghhi et al., 2011).

Inductive Category Learning

More recently, contextual interference has been explored with aural skills in music education, such as recognizing composer musical styles through listening (Wong et al., 2020). Additionally, this skill is an essential component in assessments of musicianship, especially for certificate programs like the Royal Conservatory of Music of Canada (The Royal Conservatory of Music, 2016; Wong et al., 2020). Although different musical styles are formally taught to students, the ability to recognize the styles requires inductive category learning, which is learning a concept or category after exposure to its exemplars. To do this, the learner must extract defining features from those exemplars and then use these features to distinguish stimuli from different categories (Carvalho and Goldstone, 2014).

Interestingly, interleaving appears to improve inductive category learning and aural skills in music education. In a study by Wong et al. (2020), undergraduate students

with 4 or less years of musical experience learned the musical styles of twelve composers by listening to six music excerpts from each composer. Half of the excerpts were blocked where excerpts from one composer were presented consecutively, while another half were interleaved where excerpts were randomly intermixed. After the learning phase, participants were asked to correctly identify the composer in 48 new excerpts. Despite having little musical experience, participants were significantly better at categorizing new excerpts from composers that were presented in the interleaved schedule than the blocked schedule. This supported the notion that interleaving generally improves inductive category learning (Wong et al., 2020). However, most participants believed that blocking was equally or more effective compared to interleaving, indicating that metacognitive judgment of the learning strategy was inconsistent with actual performance (Wong et al., 2020).

Currently, there are two possible explanations on how interleaving improves inductive category learning. The first is the spacing effect, which postulates that retention is enhanced when exemplars of the same category are temporally spaced apart, rather than in quick succession (Birnbaum et al., 2013). Presumably, a categorical item that is studied once requires more effort to recall when restudied after a delay (Birnbaum et al., 2013). Thus, spacing promotes more effortful retrieval of information from memory (Dunlosky et al., 2013). The second possible explanation is provided by the discriminative-contrast hypothesis. An interleaved strategy involves juxtaposing exemplars from different categories, which highlights key differences that separate one category from another (Birnbaum et al., 2013). As a result, learners become better at discriminating between categories and in the context of music, classifying exemplars of

various composers' styles. On the other hand, a blocking strategy involves fewer alternations between categories and instead emphasizes similarities within a single category or single composer's style (Birnbaum et al., 2013).

Feedback Learning

Other strategies such as feedback also play an important role in effective learning. Feedback corrects errors and reinforces mental linkages, which is an essential construct of learning theories and instructional practice (Pashler et al., 2005). In music education, students are regularly tested and provided with feedback from instructors to improve comprehension of musical concepts and technical expertise during performances (Scott, 2012). One format of feedback delivery is immediately showing the correct answer after answering a test question (Butler et al., 2007). For example, students in a classroom setting respond to a multiple-choice question and are immediately given the accuracy of response options by the instructor (Smith et al., 2019). Whereas taking a test provides an opportunity to improve long-term retention of the learning material, tests without feedback can reinforce erroneous information (Butler et al., 2007). Instead, immediate feedback reduces the discrepancy between actual and desired knowledge, allowing learners to discriminate between known and unknown concepts (Roediger and Karpicke, 2006).

Past studies demonstrating the advantage of interleaving have not concurrently investigated the effect of feedback. Although interleaving and feedback both benefit memory retention and accuracy in performance, the nature of interaction between the two learning strategies remains unclear. Additionally, studies on contextual interference

have almost exclusively focused on visual learning with images and motor skills rather than auditory learning (e.g. different styles of music).

Our Experiment

The goal of this study was to examine the effects of contextual interference and feedback on inductive category learning in the context of differentiating musical composers' styles. We hypothesized a main effect of interleaving, whereby performance for interleaved practice would be better than blocked practice, as well as a main effect of feedback, whereby performance for the feedback group would be better than the non-feedback group. When interleaving and feedback are combined as learning strategies, an additive effect was expected and would be seen through a greater increase in test performance. As an exploratory avenue, we also examined participants' metacognitive judgments of their learning and how accurately they perceived which practice strategy was more effective.

Methods

Participants

140 participants were recruited for the study. 6 participants were excluded due to missing and/or low quality data, with more specific reasons stated in the data analysis section. 134 participants (74 males, 60 females) aged between 17 and 39 years (M = 22.51, SD = 5.33) were included in the study analysis. 68 participants were classified as musicians with five or more years of musical experience. 66 participants were non-musicians with four or less years of musical experience. 27 participants were assigned to receive feedback. 107 participants were assigned to receive no feedback, who were part of a larger ongoing study in the Music and Neuroscience lab.

Participants were recruited through social media and SONA, a portal where enrolled students can take part in psychological studies conducted at Western University. Participants were granted either course credits or monetary compensation for their participation. The Western University Non-Medical Research Ethics Board (109152) approved all study procedures. All participants provided their written informed consent to participate in the study.

Materials

Music excerpts were selected from six composers: Bartok, Cage, Debussy, Ligeti, Shostakovich, and Webern. All composers were selected from the 20th century modernist era to control for differences in style arising from historical periods. Additionally, modernist music tends to be less familiar to the general population and undergraduate music students of year 2 and below, which minimizes a ceiling effect on

test scores and extraneous noise related to prior knowledge. Twenty-two music excerpts by each of the six composers were selected, totalling to 132 excerpts. Specifically, all excerpts were string quartets to ensure that participants did not classify excerpts based on instrumentation and associate composers with a certain musical ensemble configuration.

Recordings of music excerpts were obtained from an ongoing study in same lab. Each excerpt lasted 15 seconds to allow for sufficient exposure to the excerpts while maintaining participants' attentions throughout the study. Excerpts were divided across three phases: learning, acquisition, and test. Fourteen excerpts from each of the six composers were used in the learning phase, four excerpts in the acquisition phase, and the remaining four excerpts in the test phase. The study was implemented using PsychoPy v3.0, a software package written in Python programming language. The study was conducted online using Pavlovia, a platform for computerized stimulus presentation and response recording.

Procedure

The study was composed of three parts: learning phase, acquisition phase, and test phase. During the learning phase, three composers were presented in blocked schedule in which music excerpts from one composer were presented consecutively. Another three composers were presented in interleaved schedule in which music excerpts from various composers were alternated. By the end of the learning phase, each participant listened to 84 music excerpts in total, 14 from each of the six composers. The six composers were assigned to one of two groups. Group 1 comprised excerpts by Bartok, Ligeti, and Shostakovich. Group 2 comprised excerpts by Cage,

Debussy, and Webern. Order of the composer groups and practice schedules were counterbalanced, such that participants were randomly assigned to one of four counterbalanced versions of the experiment. In the first part of the learning phase, participants listened to one of the following four combinations: Group 1 blocked, Group 1 interleaved, Group 2 blocked, or Group 2 interleaved. In the second part, participants listened to the opposite format. For example, participants who listened to Group 1 blocked in the first part of the learning phase listened to Group 2 interleaved in the second part. Each excerpt was presented for 15 seconds while the respective composer's name was shown on the screen. Upon selecting the composer's name, participants were given a new excerpt. After finishing the learning phase, participants were given an optional 5-minute break before entering the acquisition phase.

During the acquisition phase at end of Day 1, participants were tested on 24 new music excerpts from the same six composers in random order. Composers' names were not shown while the excerpts were playing. Participants were asked to correctly identify the composer who wrote the excerpt by clicking on one of six names displayed on the screen. After providing a response, participants rated their level of confidence in their answer for each question on a 4-point Likert-type scale (1 = guessing, 2 = not very confident, 3 = somewhat confident, 4 = very confident). Then, participants in the feedback group were given corrective feedback (e.g. "Correct! The answer was Bartok" or "Incorrect! The answer was Ligeti"). Participants in the non-feedback group moved onto the next question without any additional information about the correctness of their answer after reporting their confidence rating.

Participants were emailed with the second part for the online study 18 hours after completing Day 1 in order to accommodate for different schedules and time zones. Then, participants were given 24 hours to complete the test phase on Day 2. Participants listened to 24 new music excerpts in random order for 15 seconds each. Similar to the acquisition phase, participants were asked to correctly identify the composer and rate their level of confidence. However, none of the participants received feedback. At the end of study, participants filled out a 27-item questionnaire, assessing sociodemographic background, musical background, and metacognitive judgments as potential predictors of test performance. For musical background specifically, participants self-reported their engagement in musical activities such as the number of hours per week practicing a musical instrument or singing, familiarity with particular musical styles, and musical training in terms of years of private instruction. Participants were debriefed about the meanings of blocked and interleaved practice. Then, they reported their metacognitive judgments of the practice schedules by selecting one of the following three outcomes: (a) blocking was better than interleaving, (b) blocking was as good as interleaving, and (c) interleaving was better than blocking.

Data Analysis

6 participants in total were excluded from the study due to missing data for Day 2. Low quality data were also excluded if time spent on each response was over 20 seconds, if participants scored less than 75% in the learning phase, and if participants completed Day 1 and Day 2 of the study more than 42 hours apart. As mentioned before, Day 2 began 18 hours after Day 1 and participants were given 24 hours to

complete the study, hence the 42 hours. For analysis, the final sample size was 134 participants.

The study used a mixed design with within-subjects and between-subjects variables. The within-subjects independent variable was practice schedule, where all participants underwent both the blocked and interleaved practice schedules. The between-subjects independent variable was the presence of corrective feedback during the acquisition phase. Each participant's test performance was measured by the percentage accuracy of classified novel music excerpts from the test phase. Mean accuracy was computed for each condition: blocked with feedback, interleaved with feedback, blocked without feedback, and interleaved without feedback.

The main analysis was conducted using a two-way mixed-design ANOVA with practice schedule (i.e. interleaved or blocked) and presence of feedback (i.e. with or without feedback) as the independent variables and accuracy in test phase on Day 2 as the dependent variable.

An exploratory analysis using a four-way mixed-design ANOVA was conducted following the main analysis in order to examine the additional effects of musical background (i.e. musician vs non-musician) and day of study (i.e. acquisition on Day 1 vs Day 2) on accuracy and long-term retention. Musical background and day of study were added as between-subjects and within-subjects independent variables, respectively.

Additionally, mean values of participants' confidence levels in their answers (Likert scales 1–4) were computed from each of the four conditions and were compared across

the conditions using two-way mixed-design ANOVA. For all analyses involving mixed-design ANOVA, a follow-up Tukey multiple comparisons test was conducted if an interaction between independent variables was significant.

Pearson's chi-square test was used to analyze participants' metacognitive judgments of practice schedules by comparing actual test performance to the frequencies of the following three outcomes: (a) blocking was better than interleaving, (b) blocking was as good as interleaving, and (c) interleaving was better than blocking. The metacognitive judgments of participants in feedback and non-feedback groups were analyzed separately. Statistical significance was determined by p values less than 0.05. All statistical analyses were conducted using RStudio 1.4.1103.

Results

Feedback and Practice Schedule on Day 2 Test Performance

There was a significant main effect of feedback ($F(1, 132) = 6.12$, $p = .015$) on mean accuracy. Participants who received feedback ($M = 30.86\%$, $SEM = 2.91$) performed significantly better in classifying musical excerpts compared to participants who did not receive feedback ($M = 22.78\%$, $SEM = 1.21$). Additionally, there was a significant main effect of practice schedule type ($F(1, 132) = 8.80$, $p = .004$) on mean accuracy. Interleaved schedule ($M = 26.18\%$, $SEM = 1.70$) showed significantly higher mean accuracy than blocked schedule ($M = 22.64\%$, $SEM = 1.53$).

However, the main effects must be interpreted with caution given that there was a significant interaction between feedback and practice schedule ($F(1, 132) = 5.30$, $p = .023$). In the feedback group, interleaving led to significantly higher mean accuracy than blocking ($p = .020$, Fig. 1). In the non-feedback group, practice schedule had no significant effect on mean accuracy ($p = .881$, Fig. 1). The interleaved with feedback condition ($M = 36.73\%$, $SEM = 4.30$) had significantly higher mean accuracy than blocked with feedback ($M = 25.00\%$, $SEM = 3.67$, $p = .020$, Fig. 1), interleaved without feedback ($M = 23.52\%$, $SEM = 1.75$, $p = .005$, Fig. 1), as well as blocked without feedback ($M = 22.04\%$, $SEM = 1.68$, $p = .001$, Fig. 1). There were no significant differences in mean accuracy amongst blocked with feedback, interleaved without feedback, and blocked without feedback ($p > .05$, Fig. 1).

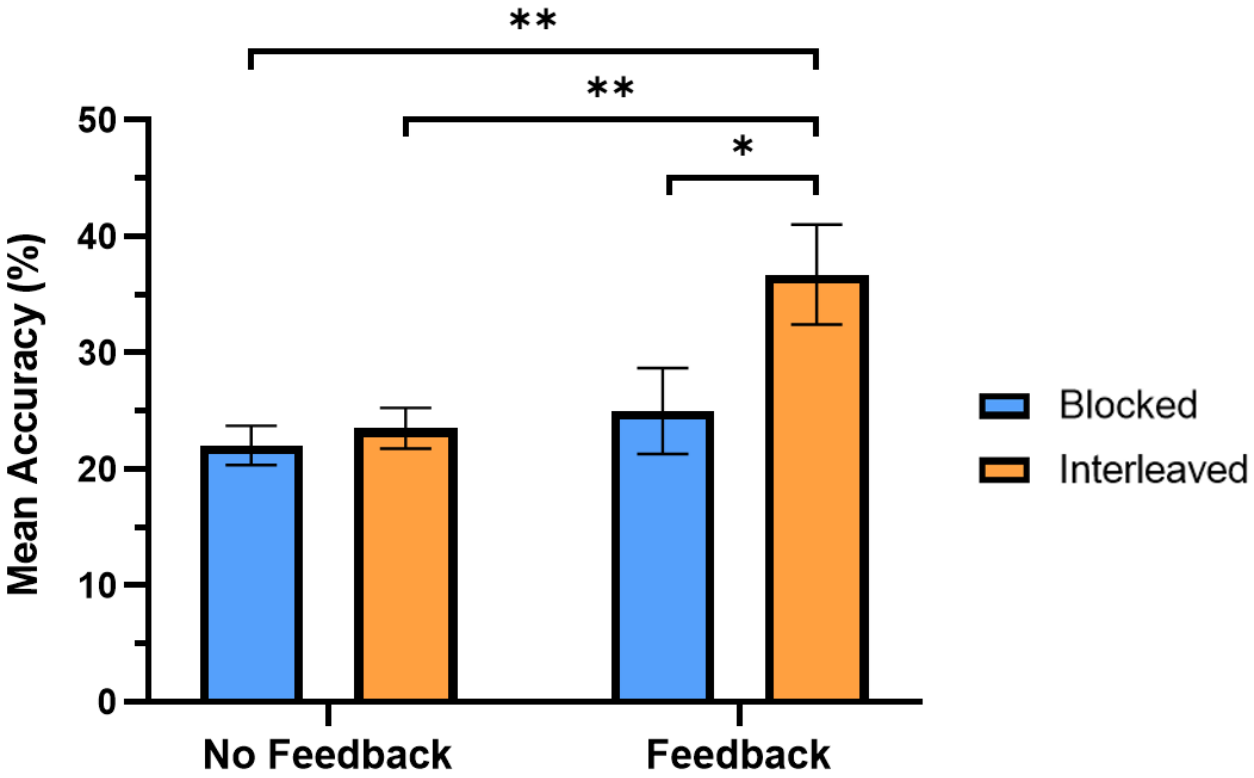

Figure 1. Mean percent accuracy of classified musical excerpts. 134 participants listened to musical excerpts from six composers, half presented in blocked schedule and another half presented in interleaved schedule. During a practice test, 27 participants were given feedback immediately after each trial. Participants were formally tested again on novel excerpts from the same six composers. Participants in feedback group showed significantly higher accuracy with interleaved scheduled compared to blocked and no feedback using a two-way mixed analysis of variance (ANOVA) with a follow-up Tukey multiple comparisons test. Data shown are mean ± SEM. * $p<0.05$, ** $p<0.01$ vs interleaved with feedback.

Day of Study and Musical Background on Test Performance

A four-way mixed ANOVA was conducted on mean accuracy that included practice schedule, presence of feedback, day of study, and musical background as independent variables. The main effect of feedback remained significant ($F(1, 132) = 9.86$, $p = .002$). However, the main effect of practice schedule ($F(1, 132) = 3.25$, $p = .074$) and interaction between feedback and practice schedule on mean accuracy were not significant ($F(1, 132) = 2.59$, $p = .110$).

There was no significant main effect of day of study ($F(1, 132) = 0.30$, $p = .584$) on mean accuracy. However, there was a significant interaction between day of study and practice schedule ($F(1, 132) = 7.36$, $p = .008$). On Day 1, there were no significant differences in mean accuracy between blocked and interleaved schedules ($p = 1.000$, Fig. 2). On Day 2, interleaved schedule ($M = 26.18\%$, $SEM = 1.70$) led to significantly higher accuracy than blocked schedule ($M = 22.64\%$, $SEM = 1.53$, $p = .015$, Fig. 2). Feedback did not significantly interact with day of study ($F(1, 132) = 1.26$, $p = .264$).

There was a significant main effect of music experience ($F(1, 130) = 20.21$, $p < .001$) on mean accuracy. Participants with 5 or more years of music experience ($M = 28.77\%$, $SEM = 1.28$) performed significantly better in classifying musical excerpts than participants with 4 or less years of music experience ($M = 19.76\%$, $SEM = .90$, $p < .001$, Fig. 3). However, this main effect must be interpreted with caution given that there was a significant interaction between music experience and feedback ($F(1, 130) = 5.17$, $p = .025$). For non-musicians, feedback had no significant effect on mean accuracy ($p = 0.909$, Fig. 3). For musicians, feedback led to significantly higher mean accuracy than no feedback ($p < .001$, Fig. 3).

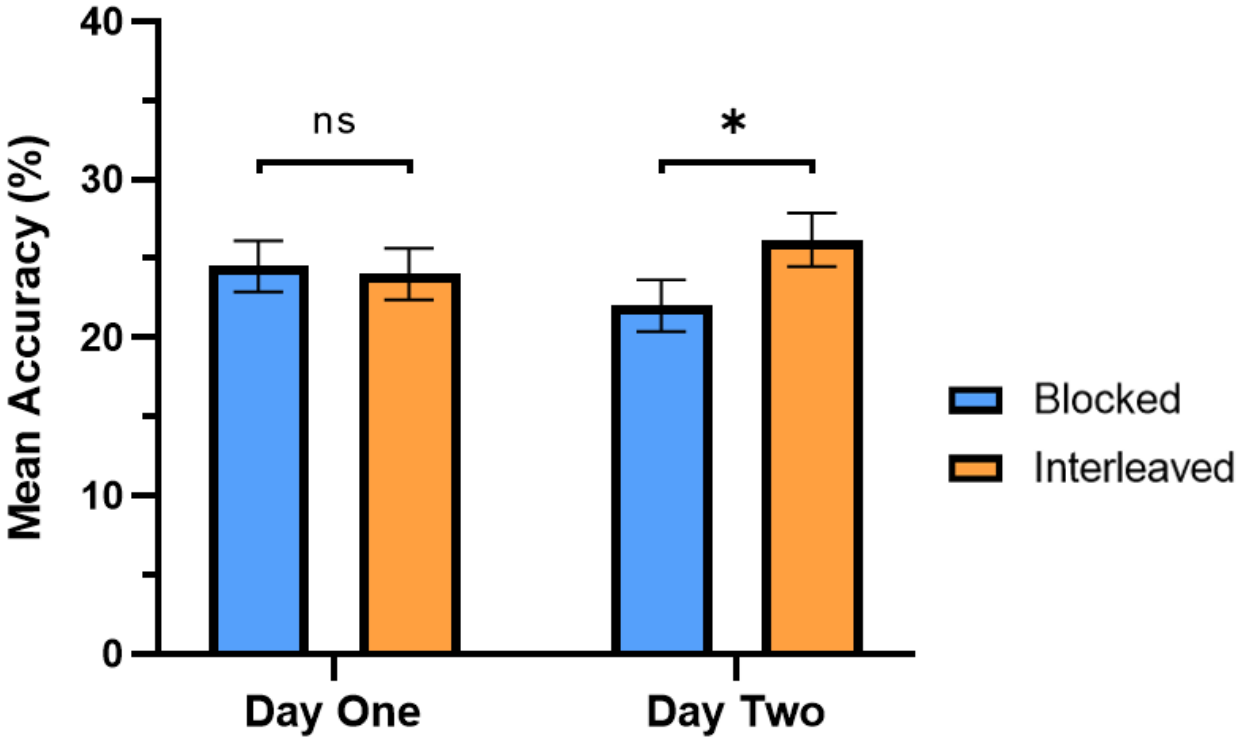

Figure 2. Mean percent accuracy of classified musical excerpts. On day one, 134 participants listened to musical excerpts from six composers, half presented in blocked schedule and another half presented in interleaved schedule. Participants practiced classifying novel excerpts from the same six composers with or without corrective feedback. On day two, participants were tested on another set of novel excerpts. On day one, there was no significant difference between blocked and interleaved schedules. On day two, interleaved schedule showed significantly higher mean accuracy compared to blocked schedule using four-way mixed analysis of variance (ANOVA) with a follow-up Tukey multiple comparisons test. Data shown are mean ± SEM. * $p<0.05$ vs interleaved on day two.

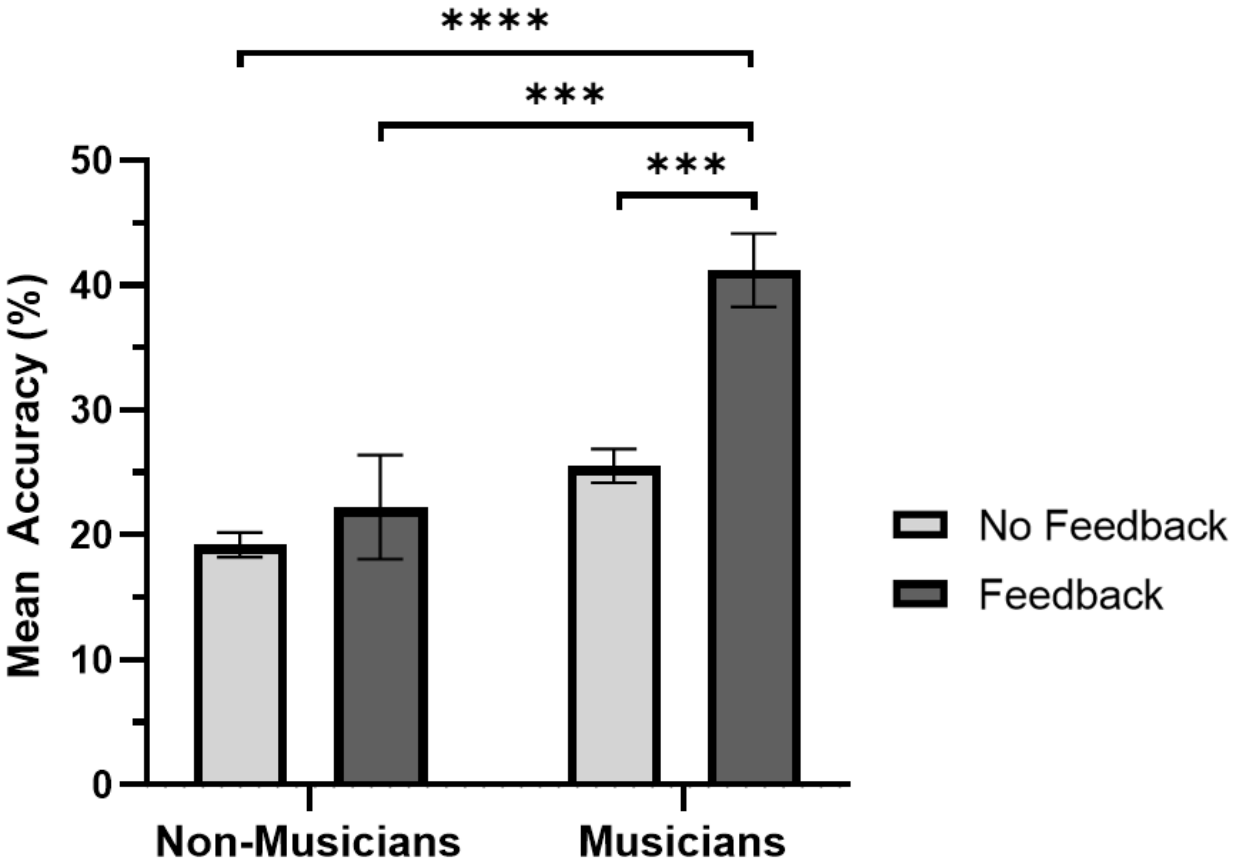

Figure 3. Mean percentage accuracy of classified musical excerpts. 134 participants (66 non-musicians, 68 musicians) listened to musical excerpts from six composers, half presented in blocked schedule and another half presented in interleaved schedule. During a practice test, 27 participants were given feedback immediately after each trial. Participants were formally tested again on novel excerpts from the same six composers. Musicians with feedback showed significantly higher mean accuracy compared to musicians without feedback and all non-musicians using four-way mixed analysis of variance (ANOVA) with a follow-up Tukey multiple comparisons test. Data shown are mean ± SEM. *** $p<0.001$, **** $p<0.0001$ vs musicians with feedback.

Musicians with feedback (M = 41.22%, SEM = 2.95) had significantly higher mean accuracy than musicians without feedback (M = 25.54%, SEM = 1.34, $p < .001$), non-musicians with feedback (M = 21.96%, SEM = 2.28, $p = .001$), as well as non-musicians without feedback (M = 19.22%, SEM = .97, $p < .001$, Fig. 3). There were no significant differences in mean accuracy amongst musicians without feedback, non-musicians with feedback, and non-musicians without feedback ($p > .05$, Fig. 3).

Feedback and Practice Schedule on Confidence Ratings

A two-way mixed ANOVA revealed a significant main effect of feedback ($F(1, 132) = 4.12$, $p = .044$) on mean confidence ratings of participants' answers. Participants who received feedback (M = 2.51, SEM = .51) showed significantly lower confidence in classifying musical excerpts compared to participants who did not receive feedback (M = 2.70, SEM = .52, Fig. 4). There was no significant main effect of practice schedule ($F(1, 132) = 0.14$, $p = .713$) nor a significant interaction between feedback and practice schedule on mean confidence ratings ($F(1, 132) = 1.25$, $p = 0.265$).

Metacognitive Judgment of Practice Schedules vs. Actual Test Performance

Overall, the majority of participants' metacognitive judgments of each practice schedule's effectiveness did not predict their actual test performance. In the non-feedback group, there was no significant association between participants' metacognitive judgments and actual test performances (X^2 (4, N = 107) = 3.168, $p = .530$). 58% (62 out of 107) of participants thought that blocking was equally or more effective compared to interleaving (Table 1). Yet, 60% (64 out of 107) of participants performed equally well or better when interleaving compared to blocking (Table 1).

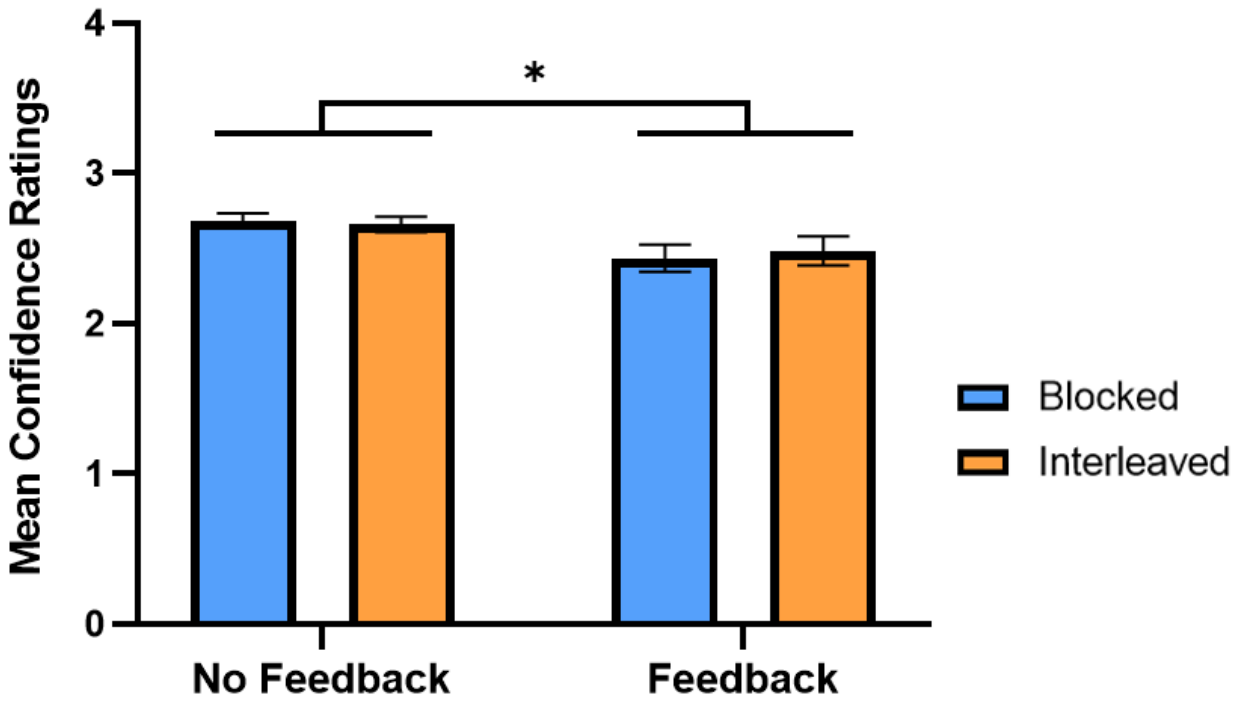

Figure 4. Mean confidence ratings of participants when classifying musical excerpts. 134 participants listened to musical excerpts from six composers, half presented in blocked schedule and another half presented in interleaved schedule. During a practice test, 27 participants were given feedback immediately after each trial. Participants were formally tested again on novel excerpts from the same six composers and gave confidence ratings for their answers. Participants in feedback group had significantly lower confidence ratings compared to no feedback group using two-way mixed analysis of variance (ANOVA) with a follow-up Tukey multiple comparisons test. Data shown are mean ± SEM. * $p<0.05$ vs no feedback.

Table 1. Frequencies of Participants' Actual Test Performance and Metacognitive Judgments of Practice Schedules in Non-Feedback Group (N=107).

		Actual Test Performance			
		Blocked > Interleaved	Blocked = Interleaved	Interleaved > Blocked	Total
Which schedule type do you think helped you learn more?	Blocked > Interleaved	21[a]	4	18	43
	Blocked = Interleaved	6	4[a]	9	19
	Interleaved > Blocked	16	8	21[a]	45
Total		43	16	48	107

[a] Participants who correctly predicted their actual test performance.

In the feedback group, there was also no significant association between participants' metacognitive judgments and actual test performance (X^2 (4, N = 27) = 5.537, p = .237). 67% (18 out of 27) of participants thought that blocking was equally or more effective compared to interleaving (Table 2). Yet, 78% (21 out of 27) of participants performed equally well or better when interleaving compared to blocking (Table 2).

Table 2. Frequencies of Participants' Actual Test Performance and Metacognitive Judgments of Practice Schedules in Feedback Group (N=27).

		Actual Test Performance			
		Blocked > Interleaved	Blocked = Interleaved	Interleaved > Blocked	Total
Which schedule type do you think helped you learn more?	Blocked > Interleaved	3[a]	2	9	14
	Blocked = Interleaved	0	0[a]	4	4
	Interleaved > Blocked	3	3	3[a]	9
Total		6	5	16	27

[a] Participants who correctly predicted their actual test performance.

Discussion

When feedback was not provided, participants' test performance in classifying novel music excerpts did not differ between composers studied in blocked and interleaved schedules. When feedback was provided, however, participants showed significantly better test performance in classifying composers studied in an interleaved schedule as opposed to blocked. The hypothesis of the study was partially supported in that the interaction between interleaving and feedback on test performance was synergistic, rather than additive. A similar study by Wong et al. (2020) also reported a significant improvement in classifying musical excerpts when composers were interleaved compared to blocked. Despite all participants receiving feedback in their study design, the improvement was largely attributed to the advantages of interleaving rather than feedback.

In contrast to Wong et al. (2020), our study did not observe an interleaving advantage when feedback was not provided. Two reasons for this observation may be the inherent difficulty of the discrimination task and the important role of feedback in correctly learning composers' musical styles (Butler and Roediger, 2008). Firstly, Wong et al. (2020) used composers from various eras for greater generalizability of their findings across many styles. To minimize ceiling effect and the effect of prior knowledge, we used composers only from the 20th century modernist era, who are generally unfamiliar up to university-level music students in year two. Composer styles also tend to be more similar within the same era than styles between eras, which inherently reduces discriminative contrast and may pose greater difficulty for novices and experts alike (Gromko, 1993; Addessi et al., 1996). Secondly, without feedback to correct the

errors, inaccurate knowledge may be reinforced while learners provide the same inaccurate answer on a later test (Butler and Roediger, 2008). In our study, participants could have failed to correctly associate the defining features from musical excerpts with a particular composer, leading to inaccurate inductive category learning and undermining the effectiveness of interleaving (Janssen et al., 2019).

That said, feedback alone was not sufficient in significantly improving the ability to differentiate composers' styles. For example, past studies have shown that auditory learning of frequency or tone discrimination tasks can occur with or without feedback (Platt and Racine, 1985; Zaltz et al., 2010). While feedback may have reduced misinformation and helped attain correct knowledge of musical styles, interleaving was also necessary for greater improvement in test performance. Alternating the order of musical categories through an interleaved schedule may have enhanced the contrast of different styles and their subsequent categorizations (Birnbaum et al., 2013). Furthermore, interleaving could have better maintained participants' cognitive engagement, leading to less mind wandering and deeper processing of the tested material (Metcalfe and Xu, 2016). Accordingly, interleaving with feedback led to better test performance, as opposed to blocking that promotes attentional lapses due to habituation and highlights similarities between categories rather than differences (Birnbaum et al., 2013).

Participants also showed greater long-term retention of composers' styles presented in interleaved schedule compared to blocked from Day 1 to Day 2. Long-term benefits of interleaving have been widely observed with inductive category learning of visual categories and other auditory-related tasks. (Zulkiply and Burt, 2013; Carter and

Grahn, 2016). Although blocking creates a sense of fluency and improvement during practice due to habituated repetition, performance immediately after learning is an inaccurate measure of long-term learning (Rendell et al., 2010). Instead, interleaving may result in more elaborate encoding processes and better memory consolidation after a retention phase that can last up to one week (Rendell et al., 2010; Zulkiply and Burt, 2013). Consequently, whether improvements persist from one day to the next may depend more on the type of processing involved and less on performance during practice.

Beyond practice strategies, musical background also affected the learning of composers' styles. Participants with 5 or more years of music experience showed significantly better test performance than non-musicians. This suggests that the ability to differentiate musical styles is influenced by more extended periods of prior training (Platt and Racine, 1985; Asmus and Harrison, 1990; Harrison et al., 1994). Furthermore, feedback led to improved performance for participants with music experience. However, previous research on the interaction between prior knowledge and feedback are inconsistent. Indeed, some studies have shown that prior knowledge may hinder learning from feedback and subsequently decrease test performance. (Krause et al., 2009; Gielen et al., 2010; Fyfe et al., 2012; Fyfe and Rittle-Johnson, 2015). On the other hand, participants with high prior knowledge may encode new information more efficiently and are thus more likely to attend to feedback when their responses are inaccurate (Sitzman et al., 2015). Although our participants with musical background benefitted more from feedback than non-musicians, the cognitive mechanism that underlies this relationship warrants further research.

Interestingly, participants' metacognitive judgments of their learning contradicted the actual benefits of feedback and interleaving. Feedback significantly lowered participants' confidence in their answers, despite leading to better actual test performance. However, people regularly overestimate their own memory performance without corrective mechanisms such as feedback (Metcalfe and Finn, 2011; Geurten and Meulemans, 2016). Overconfidence can prevent one from acknowledging the need to improve skills during practice, which may lead to poorer performance in later tests (Marteau et al., 1990). While participants may have acknowledged their poor performance when given feedback and became less confident, their metacognitive judgment was in fact more accurate (Butler et al., 2008).

Also, the majority of participants mistakenly believed that blocking was equally or more effective than interleaving, despite performing better with interleaved composers' music. These inaccurate metacognitive judgments reflect the sense of fluency that blocking elicits during practice and the belief that listening to a single composer repeatedly promotes learning of their defining characteristics (Yan et al., 2016). In an interleaved schedule, participants could have been influenced by their subjective sense of disfluency during practice, generating a perception that interleaving was less effective than blocking (Kornell and Bjork, 2008; Yan et al., 2016).

Limitations

Due to time constraints, the sample size of the feedback group (N=27) was four times smaller than the non-feedback group (N=107). The unequal sample sizes may have limited statistical power for comparing the two study groups and analyzing the effect of feedback. Nevertheless, observing significant interactions with a small sample

is promising and may indicate a sufficiently large effect size of feedback with interleaving and musical background.

Another limitation lies in the different recruitment strategies that were employed for the feedback and non-feedback groups. While most participants were undergraduate students from Western University, 75 participants in the non-feedback group were recruited from the general population using social media while all participants in the feedback group were undergraduate students. This study design may have introduced a confounding variable of sample characteristics influenced by age, education level and musical background. Therefore, the synergistic interactions between feedback and interleaving or musical background must be interpreted with caution. The study being primarily conducted with university students also limits generalizability of findings to the general population.

Future Directions

Future work should investigate why an interleaved schedule alone did not improve test performance in our study. The optimality of interleaving could be compared across different levels of task difficulty, such as varying the levels of cognitive demand while learning or using more dissimilar composers' styles than similar. Ascertaining these causal relationships in the auditory domain will require further empirical support.

The interaction between musical background and feedback should also be investigated, which was left unexplained in our current study. Past literature has consistently identified musical aptitude, academic achievement, intelligence, musical experience, and motivation for music as positive predictors of musical performance

(Asmus and Harrison, 1990; Harrison et al., 1994). Determining how these factors influence the benefits of interleaving and feedback may improve the implementation of learning strategies across diverse settings in music education.

Varying the timing and frequency of feedback delivery should also be investigated in the context of interleaving and musical learning. While our study provided only one round of feedback immediately after each question, introducing a delay before feedback presentation has been shown to increase retention of correct responses (Butler et al., 2007). Implementing multiple rounds of feedback may also provide more information for learning and improve performance (Lam et al., 2011). Changing these parameters of feedback to enhance a synergistic interaction with interleaving may be useful for optimizing learning of composers' styles.

Conclusion

Our study found that combining interleaving and feedback led to a synergistic improvement in inductive category learning of composers' styles. Despite repetition being a widely recommended strategy, music education could benefit more from cognitive research that shows the superiority of interleaving and its greater advantages when educators supplement student learning with feedback. The long-term goal of this research is to improve students' music aural skills while establishing the benefits of interleaving and feedback in the auditory domain.

Acknowledgements

I would like to thank _____ and _____ for providing me with tremendous supervision and guidance despite the challenges of an online curriculum. I would like to thank _____ for organizing all lab meetings and providing students with much-needed interactions during the pandemic. I would also like to thank all members of the Grahn lab for contributing to a wonderful experience with research in music and neuroscience.

References

Addessi A, Baroni M, Luzzi C, and Tafuri J (1996) The development of musical stylistic competence in children. *Bull Counc Res Music Educ* **127**:8–15.

Asmus EP, and Harrison CS (1990) Characteristics of motivation for music and musical aptitude of undergraduate nonmusic majors. *J Res Music Educ* **38**:258–268.

Austin JR, and Berg MH (2006) Exploring music practice among sixth-grade band and orchestra students. *Psychol Music* **34**:535–558.

Battig WF (1966) Acquisition of skill, in *Facilitation and interference* (Bilodeau EA ed) pp 215–244, Academic Press, New York.

Birnbaum MS, Kornell N, Bjork EL, and Bjork RA (2013) Why interleaving enhances inductive learning: the roles of discrimination and retrieval. *Mem Cogn* **41**:392–402.

Bjork RA (1994) Memory and metamemory considerations in the training of human beings, in *Metacognition: knowing about knowing* (Metcalfe J and Shimamura AP eds) pp 185–205, The MIT Press, Cambridge.

Butler AC, Karpicke JD, and Roediger HL (2007) The effect of type and timing of feedback on learning from multiple-choice tests. *J Exp Psychol Appl* **13**:273–281.

Butler AC, Karpicke JD, and Roediger HL (2008) Correcting a metacognitive error: feedback increases retention of low-confidence correct responses. *J Exp Psychol Learn Mem Cogn* **34**:918-928.

Butler AC, and Roediger HL (2008) Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Mem Cognit* **36**:604–616.

Carter CE, and Grahn JA (2016) Optimizing music learning: exploring how blocked and interleaved practice schedules affect advanced performance. *Front Psychol* **7**:1251, Frontiers.

Carvalho PF, and Goldstone RL (2014) Effects of interleaved and blocked study on delayed test of category learning generalization. *Front Psychol* **5**:1–11.

Dunlosky J, Rawson KA, Marsh EJ, Nathan MJ, and Willingham DT (2013) Improving students' learning with effective learning techniques: promising directions from cognitive and educational psychology. *Psychol Sci Public Interes Suppl* **14**:4–58.

Feghhi I, Abdoli B, and Valizadeh R (2011) Compare contextual interference effect and practice specificity in learning basketball free throw. *Procedia Soc Behav Sci* **15**:2176–2180.

Fyfe E, and Rittle-Johnson B (2015) Feedback both helps and hinders learning: the causal role of prior knowledge. *J Educ Psychol* **108**:82–97.

Fyfe E, Rittle-Johnson B, and DeCaro M (2012) The effects of feedback during exploratory mathematics problem solving: Prior knowledge matters. *J Educ Psychol* **104**:1094–1108.

Geurten M, and Meulemans T (2016) The effect of feedback on children's metacognitive judgments: a heuristic account. *J Cogn Psychol* **29**:184–201.

Gielen S, Peeters E, Dochy F, Onghena P, and Struyven K (2010) Improving the effectiveness of peer feedback for learning. *Learn Instr* **20**:304–315.

Goode S, and Magill RA (1986) Contextual interference effects in learning three badminton serves. *Res Q Exerc Sport* **57**:308–314.

Gromko JE (1993) Perceptual differences between expert and novice music listeners: A multidimensional scaling analysis. *Psychol Music* **21**:34–47.

Hall KG, Domingues DA, and Cavazos R (1994) Contextual interference effects with skilled baseball players. *Percept Mot Skills* **78**:835–841.

Harrison CS, Asmus EP, and Serpe RT (1994) Effects of musical aptitude, academic ability, music experience, and motivation on aural skills. *J Res Music Educ* **42**:131–144.

Hoppmann RA, and Patrone NA (1989) A review of musculoskeletal problems in instrumental musicians. *Semin Arthritis Rheum* **19**:117–126.

Janssen CP, Everaert E, Hendriksen HMA, Mensing GL, Tigchelaar LJ, and Nunner H (2019) The influence of rewards on (sub-)optimal interleaving. *PLoS One* **14**:e0214027, Public Library of Science.

Kornell N, and Bjork R (2008) Learning concepts and categories is spacing the "enemy of induction"? *Psychol Sci* **19**:585–592.

Krause U-M, Stark R, and Mandl H (2009) The effects of cooperative learning and feedback on e-learning in statistics. *Learn Instr* **19**:158–170.

Lam CF, DeRue DS, Karam EP, and Hollenbeck JR (2011) The impact of feedback frequency on learning and task performance: challenging the "more is better" assumption. *Organ Behav Hum Decis Process* **116**:217–228.

Magill RA, and Hall KG (1990) A review of the contextual interference effect in motor skill acquisition. *Hum Mov Sci* **9**:241–289.

Marteau TM, Wynne G, Kaye W, and Evans TR (1990) Resuscitation: experience without feedback increases confidence but not skill. *BMJ* **300**:849–850.

Maynard LM (2006) The role of repetition in the practice sessions of artist teachers and their students. *Bull Counc Res Music Educ* **167**:61–72.

Metcalfe J, and Finn B (2011) People's hypercorrection of high-confidence errors: did they know it all along? *J Exp Psychol Learn Mem Cogn* **37**:437–448.

Metcalfe J, and Xu J (2016) People mind wander more during massed than spaced inductive learning. *J Exp Psychol Learn Mem Cogn* **42**:978–984.

Pashler H, Cepeda NJ, Wixted JT, and Rohrer D (2005) When does feedback facilitate learning of words? *J Exp Psychol Learn Mem Cogn* **31**:3–8.

Platt JR, and Racine RJ (1985) Effect of frequency, timbre, experience, and feedback on musical tuning skills. *Percept Psychophys* **38**:543–553.

Rendell MA, Masters RSW, Farrow D, and Morris T (2010) An implicit basis for the retention benefits of random practice*. J Mot Behav* **43**:1–13.

Roediger HL, and Karpicke JD (2006) The power of testing memory: basic research and implications for educational practice. *Perspect Psychol Sci* **1**:181–210.

Scott SJ (2012) Rethinking the roles of assessment in music education. *Music Educ J* **98**:31–35.

Shea JB, and Morgan RL (1979) Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *J Exp Psychol Hum Learn Mem* **5**:179–187.

Simon DA (2007) Contextual interference effects with two tasks. *Percept Mot Skills* **105**:177–183.

Sitzman D, Rhodes M, Tauber S, and Liceralde V (2015) The role of prior knowledge in error correction for younger and older adults. *Aging Neuropsychol Cogn* **22**:1–15.

Smith AC, Ralph BCW, MacLeod CM, and Smilek D (2019) Test feedback and learning: student preferences and perceived influence*. Scholarsh Teach Learn Psychol* **5**:255–264.

Wong SSH, Low ACM, Kang SHK, and Lim SWH (2020) Learning music composers' styles: to block or to interleave? *J Res Music Educ* **68**:156–174.

Yan V, Bjotk E, and Bjork R (2016) On the difficulty of mending metacognitive illusions: a priori theories, fluency effects, and misattributions of the interleaving benefit. *J Exp Psychol Gen* **145**:918–933.

Zaltz Y, Roth DA-E, and Kishon-Rabin L (2010) Does feedback matter in an auditory frequency discrimination learning task? *J Basic Clin Physiol Pharmacol* **21**:241–254.

Zulkiply N, and Burt JS (2013) Inductive learning: does interleaving exemplars affect long-term retention? *Malaysian J Learn Instr* **10**:133–155.