Modeling and Optimizing Forgetting and Spacing Effects during Musical Interval Training

Philip I. Pavlik, Jr.
University of Memphis
Psychology Dept.
Memphis, TN 38152
1-901-678-2326
ppavlik
@memphis.edu

Henry Hua
University of Memphis
Psychology Dept.
Memphis, TN 38152
1-901-678-5590
hyhua
@memphis.edu

Jamal Williams
University of Memphis
Psychology Dept.
Memphis, TN 38152
1-901-678-2364
jawllm10
@memphis.edu

Gavin M. Bidelman
University of Memphis,
Sch. Comm. Sci. & Disorders
Memphis, TN 38152
1-901-678-5826
gmbdlman
@memphis.edu

ABSTRACT

From novice to expert, almost every musician must recognize musical intervals, the perceived pitch difference between two notes, but there have not been many empirical attempts to discover an optimal teaching technique. The current study created a method for teaching identification of consonant and dissonant tone pairs. At posttest, participants increased their ability to discern tritones from octaves, and performance was better for those who received an interleaving order of the practice trials. Data mining of the results used a novel method to capture curvilinear forgetting and spacing effects in the data and allowed a deeper analysis of the pedagogical implications of our task that revealed richer information than would have been revealed by the pretest-to-posttest comparison alone. Implications for musical education, generalization learning, and future research are discussed.

Keywords

Model-based discovery, forgetting, interleaving, spacing effect, computer adaptive training, musical consonance

1. INTRODUCTION

Music is a rich multi-modal experience which taps a range of both perceptual and cognitive mechanisms. As in other important facets of human cognition (e.g., speech/language), music consists of constituent elements (i.e., scale tones) that can be arranged in a combinatorial manner to yield high-order units with specific categorical labels (e.g., intervals, chords). In Western tonal music, the octave is divided into 12 pitch classes (i.e., semitones). When combined, these pitch relationships can be used to construct twelve different chromatic intervals (each one semitone above or below another) that are labeled according to the relationship between fundamental frequencies of their tones. For example, two tones can form an octave (2:1 ratio), a perfect fifth (3:2 ratio), or a variety of other tonal combinations.

Perceptually, musical intervals are typically described as either consonant, associated with pleasantness and smoothness (e.g., octave), or as dissonant, associated with unpleasantness and roughness (tritone, ratio: 11:8). Helmholtz (1895) defined this "roughness" as the fluctuations in amplitude perceived by a listener which occurs when the distance between partials is small enough for them to interact (i.e., "beat") within the auditory periphery. Consonance, on the other hand, occurs in the absence of such beating, when low-order harmonics are spaced sufficiently far apart to not interact (e.g., octave, perfect fifth). Behavioral studies demonstrate that listeners treat the various intervals hierarchically and tend to prefer consonant over dissonant pitch relationships [2, 7, 10]. It is this hierarchical arrangement of pitch which largely contributes to the sense of a musical key and harmonic structure in tonal music [8].

The ability to discriminate consonant and dissonant intervals is far better than chance [20]. Indeed, this ability emerges early in life as both newborns and infants show a robust preference for consonant over dissonant tone pairs, well before being exposed to the stylistic norms of culturally specific music [5]. As such, it is posited that the perceptual distinction between consonance and dissonance might be rooted in innate auditory processing [1, 2]. While the perceptual *discrimination* of music intervals is fairly well studied [1, 2, 7, 9] the ability to *identify* intervals remains poorly understood. While the capacity to distinguish aspects of musical structure might be present at birth, the orientation towards culture-specific music and its definitions must progress during childhood. This developmental perspective of musical learning marks the vital periods for receptivity of musical structures. However, musical training is not a standard thus resulting in a continuum of musical ability amongst the general population.

Here, we ask how individuals learn to *identify* (i.e., categorically label) the pitch combinations of music. Interestingly, musically naïve individuals could potentially benefit from musical training. Research suggests that late musical training might compensate for some gaps in musical knowledge such as the enhancement of automatic encoding of interval structures [4]. Additionally, music training is said to have both short-term and long-term benefits in regards to transfer [19]. Indeed, the benefits of prolonged musical training could have long-term effects such as improved executive functioning and perceptual organization [5]. Specifically, studies suggest that training greatly influences the developing brain, making music training a promising model for examining learning [12, 13]. If music training is a potential model for investigating learning then the implementation of different training regimens should help elucidate the general learning process.

Specifically, by training individuals to identify different *harmonic intervals*, a more optimal training regimen could be derived and contribute to our understanding of pedagogical strategies. (A harmonic interval is when two tones are played simultaneously, as opposed to a melodic interval where two tones are sequential.) As previously stated, studies have shown that people already have some prior knowledge when it comes to interval discrimination, regardless of enculturation effects or innate capacities. Indeed, listeners appear to be better at distinguishing some intervals over others. For example, intervals of small integer ratios, such as perfect fourths (4:3) and perfect fifths (3:2), are typically more difficult to discriminate and are often confused [6, 18]. Furthermore, the order in which people learn certain intervals has been seen to have an effect on their ability to make accurate distinctions, e.g. [6]. If factors such as these could be understood more deeply, it may lead to more effective ways of configuring musical training.

2. METHOD

2.1 Participants

After screening our data for participants (Amazon Turk workers) who provided a full set of responses, without omissions, we had 220 participants. The average participant age was 29.88 years (*SD* = 10.32). Participants had a mean of 4.30 years of musical experience (*SD* = 5.70). Parsed into different types of musical training, participants had on average 1.34 years (*SD* = 2.31) of training with a private tutor, 2.27 years (*SD* = 4.78) studying music on their own, and 2.86 years (*SD* = 3.38) of training in a formal school setting. Prior ear training experience averaged 0.47 years (*SD* = 1.61), 0.24 years (*SD* = 0.81) of which focused specifically on harmonic interval training.

2.2 Procedure

Participants self-selected this study from a list of various available research participation opportunities on Amazon Mechanical Turk, an online data-collection service. Participants were paid $3. Participants began the study with a survey. Items included demographic information such as age and sex. This survey also asked for various predictors, including years of different types of musical training (overall, private tutoring, school), and types of musical training (ear training, harmony, reading music).

Upon finishing the survey, participants completed the interval identification task. Prior to the pretest, to ensure that people understood the task, the following instructions were given for participants to view (participants clicked a button after reading):

> Hint: The octave interval is sometimes described as smooth, pleasing or pure. The tritone interval is sometimes described as harsh, diabolic or impure.
>
> Task: Please listen to each 2 second interval, then type 'o' for octave or 't' for tritone. After each incorrect response, you are provided review to help you learn.
>
> Goal: Practice the sound identification task, attempting to learn the interval (octave or tritone) between two notes played at the same time. This pretest portion will get an initial measure of your skill in the task, and will be followed by 96 training practices, and finally a posttest of 32 practices.

This task contained three stages: practice, learning, and posttest. The practice section presented 32 intervals for the participant to label as either a tritone or octave. In the learning section, 96 intervals were presented and participants were asked to similarly label the intervals as tritone or octave. The orders of the learning intervals were presented in various sequences (see section 2.3, Conditions). Sequence type was thus the primary independent variable of the study. The posttest section, which is the primary dependent variable of the study, once again presented 32 intervals to be labeled as tritone or octave.

All trials presented the interval sound file which lasted 2 seconds and then they typed ‘t’ or ‘o’ to indicate their response (trials timed-out after 2 minutes, but learners typically responded in less than 2 seconds). After responding, a checkmark appeared for .5 seconds to indicate the selected answer was correct; if incorrect, the correct answer was given with a replay of the now labeled interval sound just responded to incorrectly, these “study” opportunities lasted 5 seconds (so there was 3 seconds of silence after each replay).

2.3 Conditions

This experiment varied the presentation sequencing to test the effectiveness of various presentation orders on the task of identifying tritone and octave harmonies. Practice trials were presented in combinations of progressive and interleaving orders organized into four blocks, each containing 24 harmonic intervals. Conditions were randomized between subjects.

A progressive order presented the harmonic intervals in consecutive blocks, each block containing the same two intervals but presented at a higher pitch register than the previous block. Block 1 contained intervals in a low register (155.6Hz/311.1Hz for octave and 185Hz/261.6Hz for tritone) block 2, intervals of a medium-low register (277.2Hz/554.4Hz for octave and 329.6Hz/,466.2Hz for tritone), block 3, intervals from a medium-high register (493.9Hz/987.8Hz for octave and 587.3Hz/830.6Hz for tritone) and block 4, intervals from a high register (880Hz/1760hz for octave and 1046.5Hz/1480Hz for tritone). Sounds were synthesized by MIDI using instrument 1 (Piano) for a 2-second duration. An antiprogressive order presented harmonic intervals in a way that made each block maximally different from the previous block. block 1 consisted of low register tones, block 2, high register tones, block 3 medium-low tones, and block 4, medium-high tones.

An interleaving order introduced a new register for each of blocks 2-4 according to the antiprogressive or progressive order, with tones already heard from the previous blocks interleaved with the new material. In other words, new registers were taught while practicing the old ones, with an equal distribution for each of the presented tone levels within a block of 24. Conditions lacking an interleaving order did not repeat tones from previous blocks.

Therefore, the 4 experimental conditions contained all 4 combinations of progressive and interleaving orders: progressive and no interleaving, antiprogressive and no interleaving, progressive with interleaving, and antiprogressive with interleaving. As a control group, there was one condition that presented 96 learning in 4 blocks that were fully mixed, (just like the pretest and posttest). For all conditions, although each block contained a predetermined set of tones, tones were randomized within each block. Finally, practice during learning “blocks” were not marked by a brief pause with an introduction screen like the pretest, learning, and post-test were marked. In other words, transitions between sets of different items during practice were not signaled to subjects.

3. RESULTS

An analysis was conducted on the number of participants in each of the five conditions (M = 44.00, SD = 5.10), to assess whether dropout was more prevalent in certain conditions. The control condition had the fewest observations, with only 37 participants, whereas the largest group, the interleaving–progressive condition, had 51 participants. Since the control condition was most difficult, we thought that might be causing this disparity. However, attrition did not differ statistically between conditions, $\chi 2(68) = 78.15$, $p = .2$.

Means and standard deviations by condition for the post-test were as follows: Condition 1 (progressive only): Mean=.78, SD=.15; Condition 2 (progressive + interleaving): Mean=.82, SD=.16; Condition 3 (no progressive or interleaving): Mean=.79, SD=.14; and Condition 4 (interleave only): Mean=.86, SD=.16. There was statistically significant improvements, averaged across conditions, from pretest to posttest, $t(219) = 9.75$, $p < .01$. Upon finding an overall positive effect from pre to posttest, analyses focused on systematic differences between interleaving and progressive tone presentations in the practice trials. Figure 1 shows average performance across the entire 160 trials of the experiment, averaged in blocks of 8 trials. The four experimental conditions (progressive without interleaving, antiprogressive with interleaving, progressive–interleaving, and antiprogressive without interleaving) were analyzed according to two dichotomous criteria: those with or without progressive presentations, and those with or without interleaving presentations. The control condition, which presented intervals in a random order, was not used in this 2 way (interleaving/no interleaving) x 2 (progressive/antiprogressive) analysis of variance (though the results in that condition are in the same direction as below, since the highly interleaved control also performed well for learning gain relative to the blocked conditions). We used pretest score as a covariate.

The ANCOVA revealed an effect of interleaving during practice trials on posttest scores, even when controlling for pretest scores. Interleaving trial presentation order had a significant positive effect on posttest scores, $F(1, 178) = 5.81$, $p = .02$, $d = .34$. Progressive and antiprogressive ordering had no reliable effect on posttest scores, [$F(1, 178) = .48$, $p = .5$]. The interaction of progressive and interleaving orders was also non-significant, $F(1, 178) = .33$, $p = .6$.

The last series of analyses looked for any relationships between individual differences in musical training or skill with posttest scores and the overall magnitude of the pretest–posttest improvement. A list of correlations is listed in Table 1. There were modest relationships between various musical training-related and skill-related predictors and posttest scores, and some suggestion of a negative relationship between previous learning and improvement.

Overall, results suggest that the teaching paradigm created for this experiment caused improvement from pretest to posttest. This corroborates previous work [6], which presented the practice trials in an interleaved rather than sequential order, also showing a reliable advantage. Certain qualities of musical skill and training had modest relationships with posttest performance, but were negatively related to improvement. This negative correlation suggests that proficient musicians had less to gain than poorly skilled musical learners from our practice procedure.

Table 1. Correlations between participant factors and scores.

Participant factor	Posttest	Improvement
Years overall musical training	.28*	-.10
Years of musical tutoring	.25*	-.10
Ability to read music	.26*	-.16*
Understanding of musical harmonies	.30*	-.15*
Ability to hear musical harmonies	.30*	-.10
Years of self-directed musical training	.26*	-.08
Years of musical training in school	.31*	-.07
Years of ear training	.13	.03
Years of ear training on harmonies	.18*	.04

Note. N = 220, *p < .05

4. MODEL BASED DISCOVERY

While this paper mines data from a novel musical educational task, the computational model of the data we created is based on many of the common principles of educational data mining. To begin, the model builds upon a simple additive factors model (AFM) [3, 21]. The AFM is a logistic regression model that is based on the logic of counting prior practice events, so that the prior practice is an "additive factor" predicting future performance. In this early stage of research to understand how people learn musical intervals, we began with the simple assumption that each of our 8 stimuli was a knowledge component that could be learned independently in some sense. In a later model in this paper, we will also demonstrate how we can add a generalization component specific to each interval, but our current experimental design was not appropriate for more deep analysis of generalization primarily because the conditions only used 2 intervals and did not vary the spacing of each interval, rather, in all conditions, there was always a 50/50 chance of either interval for each trial.

However, while we fit a model that is built upon AFM principles of counting prior instances of practice of particular types, we found that the simple AFM model could not account for some of the effects in Figure 1. The first effect that could not be captured by AFM was simple interference based forgetting as a function of the number of trials since the prior repetition. We can see this forgetting effect (probably driven by multiple processes, including interference) manifesting in the practice block differences and transitions between practice blocks. For example, note how fast learning is in each block when items are blocked, but then observe the huge decrement when items are mixed in the final posttest. Retention is apparently quite poor when there is other practice between repetitions of the same type.

We can see similar differences comparing performance between blocks also, noting that as items are added in the interleaved conditions performance steps down with each additional register added to the set of stimuli blocked together. The mechanism behind this forgetting effect is not entirely clear. However, previous work demonstrates that the perceptual distinction between musical pitch relationships is continually strengthened with exposure and training [11]. Thus, it is possible that periodic lapses in performance across the practice blocks might be due to the fact that our non-musician listeners' internal templates for the intervals are not yet robust, rendering the mapping between interval sound and label unstable, and ultimately hindering behavioral identification.

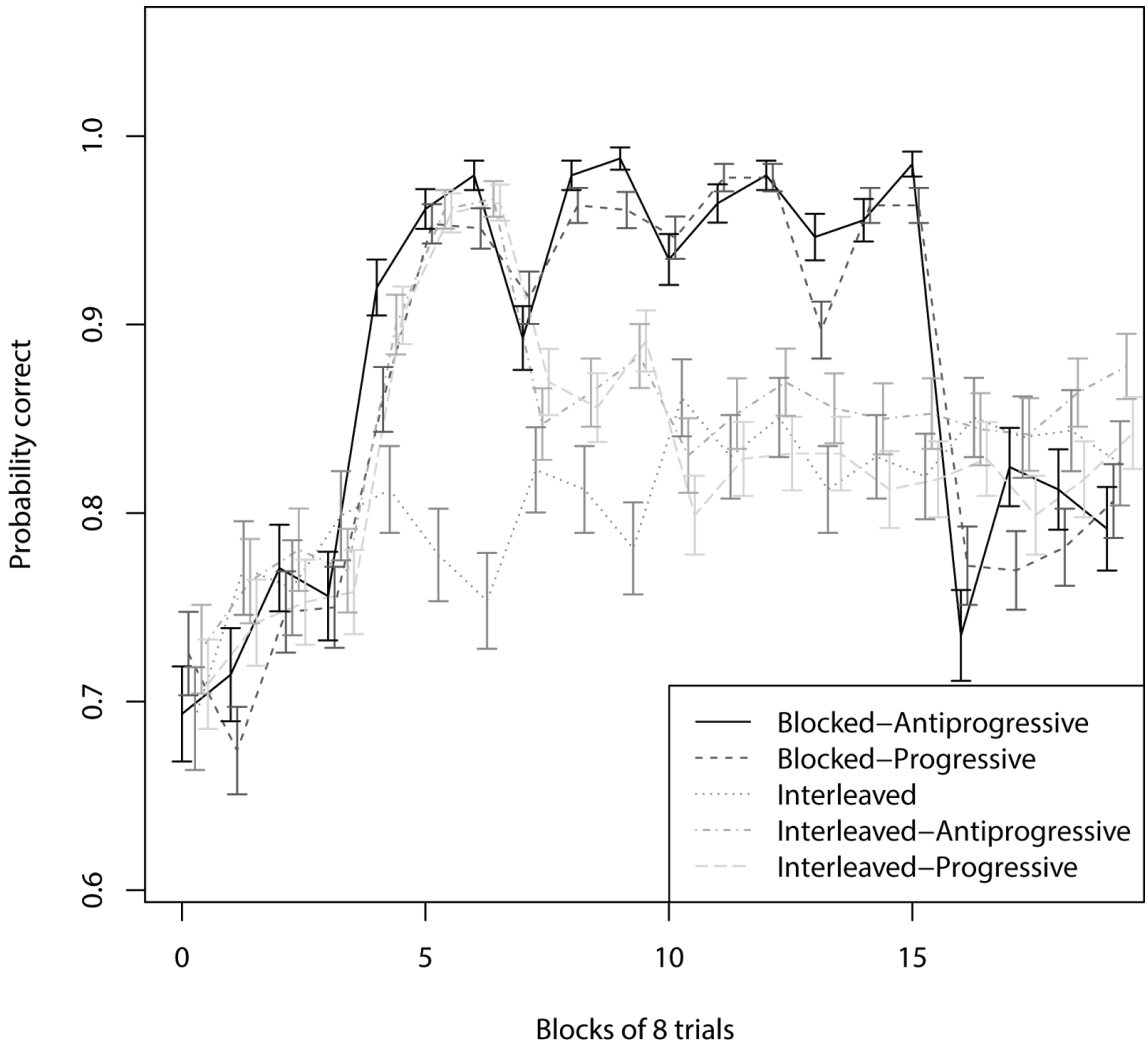

Figure 1. Performance in all conditions plotted in blocks of 8 trials. Error bars are 1 *SE* confidence intervals. First 4 and last 4 blocks represent pre- and post-test trials.

The second effect that AFM was unable to capture was the benefit of spacing/interleaving that we saw in the interleaved conditions with the ANOVA analysis. This was described in Section 3 as a significant benefit for the 2 conditions which employed more interleaving. We can also see this effect in Figure 1 as a visible difference between the interleaving and blocked conditions at posttest. This "spacing" effect, which is also very common in verbal memory experiments [16], was not by itself as strong as the effect of forgetting, but it has important implications for education. If musical educators can make use of this effect, our data suggest they may enhance learning. While AFM does not capture such an effect in the original incarnation, more complex models can capture such data. For example, Pavlik and Anderson [16] describe one such model, which functions by proposing less decay for as a function of the increased difficulty of more widely spaced practice. Unfortunately, such models have several parameters and no analytic method of solution, so, solving these models is extremely difficult due to issues of time and local minima. To resolve some of these issues we wanted to find a way to fit a similar model as Pavlik and Anderson, that relied less on an ad hoc, difficult to solve (albeit accurate) model form and more on an established model formalism (logistic regression).

The problem was that both of the effects we wanted to model, forgetting and spacing, depend on a model where each observation is predicted by a parameterized function of prior knowledge (not just the count, as in AFM), and typical logistic formalisms only allow the independent data for each observation to be used to predict the dependent effect. Of course, AFM solves this by keeping a simple count of prior practices for a skill or item-type and adding it to the data for each row so that these prior events can have an effect. This will not work easily for decay however, unless we want to have learning decrease linearly when other items are practiced. However, linear forgetting has never been considered a viable model [17]. To work around this limitation, we decided to model decay by adding a parameter that captures a percentage loss for each item as each other item is practice (exponential decay).

So, for example, if we look at the vector of prior events for some arbitrary item we might notice it is practiced at opportunity 3 and opportunity 7. Normally, in AFM this would be represented as the vector 0, 0, 0, 1, 1, 1, 1, 2, 2, 2..., thus the student would have 1 prior counted for the prediction of how they will perform on trial

7, and if they performed on trial 10, they would be credited 2 priors when computing their chance. Note, how the credit is lagged so that when we compute performance for trial 3 we do not (quite sensibly) have any priors, but when we compute the prediction for trial 4, we then have 1 prior.

With the new mechanism we introduce a "forgetting rate", *d*, that we estimated and applied to computing the prior credit vector. This decay was applied to each prior practice independently, so that if decay was say .7 (for our example above) we would have the vector 0, 0, 0, .7, .49, .34, .24, 0.17, 0.12, 0.08, and the vector 0, 0, 0, 0, 0, 0, 0, .7, .49, .34 summed equals 0,0, 0, .7, .49, .34, .24, .87, .61, .43. So, given the .7 parameter value, the student would have .24 prior counted when they perform on trial 7, and if they performed on trial 10, they would be credited .43 prior decayed strength when computing their chance.

Next, we wanted to add spacing effects to the model by using the ideas from Pavlik and Anderson [16]. In this work, Pavlik and Anderson proposed that long-term learning benefit (decayed remnant) for spaced practice was an inverse function of the current strength. To adapt this model we used the decaying strength vector as an exponent in a power function model where we estimated the base as another new parameter, *g*. So, for example, given a spacing parameter g=.005, we find that .005^0 = 1 long term learning (e.g. for the first trial, which is the 3rd opportunity in our example above) while .005^.24 = .28 long-term learning for the 7th opportunity. Long-term learning is a new vector that works in addition to the decay strength to predict performance. In our example, we would sum 0 , 0, 0, 1, 1, 1, 1, 1, 1, 1 for the first practice and 0, 0, 0, 0, 0, 0, 0 , .28, .28, .28 for the second practice. This long-term learning is permanent.

This model was estimated by nesting a logistic general linear model (GLM) within a general gradient descent optimization function. This wrapper optimization took the decay and spacing parameters, transformed the data vectors based on those parameters, and then computed the optimal logistic model and outputted the fit of that model to the wrapper. The wrapper then used the internal model's fit to adjust spacing and decay by brute force gradient descent steps (the bounded BFGS method from the `optim` function, built into R), to get a global optimization for the wrapped GLM function given the decay and spacing parameters. Figure 2 shows this optimization structure in R code where `temp` is a vector that holds the decay and spacing parameters.

```
model <- function(temp, data) {
   compute data
      as a function of temp
   compute GLM model fit using data
   return log likelihood fit}

optim(temp, model, method
   =c("L-BFGS-B"),lower=0, upper=1))
```

Figure 2. Wrapper optimization loop pseudocode.

The GLM model included fixed effects to capture the 2x4 main effects and interactions caused by the particular tones and intervals, and the prior decayed strength and prior long-term learning for the particular stimuli. Figure 3 shows the GLM structure (i.e. the independent variables that predict the dependent), which shows how we fit a single coefficient for the effect of prior decayed strengths, and a single coefficient for the effect of the long-term benefits (using the `I` function in R allows us to use the vector sum since each vector applies independently of the prediction). This means that the data vectors for these values were linearly scaled inside the GLM, while being created in the wrapper. This allows us to fit a much more complex model than if we just used the wrapper, since brute force gradient descent would have been prohibitively slow with 3 (or more) parameters. Instead, putting the GLM in the wrapper allows us to fit the minimum number of non-linear parameters (2) inside the slow brute force procedure, and then optimize several more parameters in the efficient GLM logistic function. Table 3 shows the more complex AFM-decay-space model compared to two simpler models via cross validation. The R code for the model equation first finds a parameter for the 8 decaying vectors for the 8 components (octave0, etc.). Fitting a single parameter for the effect of each of the 8 vectors simplifies the model under that assumption that forgetting is equivalent for each register by interval combination (using the `I` function in R allows us to sum vectors since each vector applies independently of the prediction). Similarly, we also assume a single parameter for the permanent learning vectors (soctave0, etc.), which account for the long-term learning from spaced practice for each stimulus type. Finally, the interval by tone interaction captures fixed-effect differences that may be due to average effects of poor fidelity of the participants' audio speakers or hearing in some registers and any other specific differences in the baseline performance with each register by interval pair. However, it might be noted that the variety of significant differences for tone and interval were not well controlled for (e.g. order of introduction) in our design, so we choose not to analyze them here.

```
answer ~

I(octave0 + octave10 + octave20 + octave30 +
tritone0 + tritone10 + tritone20 +
tritone30) +

I(soctave0 +
    soctave10 + soctave20 + soctave30 +
stritone0 + stritone10 + stritone20 +
stritone30) +

interval * tone
```

Figure 3. Logistic GLM model structure.

We tested 3 models with 5 runs of 10 fold cross validation to confirm the model generalized to our data as shown in Table 2. The three models were AFM, which simply summed the prior practices in for each of the 8 stimuli, AFM-decay, which included the new decay mechanism, and AFM-decay-space, which further layered the spacing effect mechanism into the model and is shown in Figure 3. While the test analysis reveals strong significant difference for the AFM-decay model for both *r* and *MAD*, the AFM-decay-spacing model was barely (Z=2.16, p<.05) better in terms of *r* and not significant for MAD. We added spacing effects after decay effects, since spacing effects are very small compared to decay effects in an experiment with only one-session, and thus a spacing effect mechanism may inappropriately capture effects due to decay unless decay effects are removed first (the spacing mechanism may actually improve the model in such a case, but parameter values will not meaningfully indicate a benefit to spacing, perhaps even the inverse).

Table 2. Cross validation results.

Model	Train		Test	
	Spearman *r (SE)*	*MAD (SE)*	Spearman *r (SE)*	*MAD (SE)*
AFM	0.10974 (0.00047)	0.25315 (0.0005)	0.1067 (0.00392)	0.25347 (0.00203)
AFM-decay	0.24013 (0.00044)	0.24167 (0.00053)	0.23783 (0.00402)	0.24206 (0.00239)
AFM-decay-space	0.25211 (0.00037)	0.23987 (0.00042)	0.24921 (0.0034)	0.24026 (0.00197)

It was also interesting to check how well the model could be used to simulate the experiment. Figure 4 shows graphically how well the model captures the aggregate effects. Note that even the error bars are of very similar magnitude. This simulation was constructed by generating random number from 0 to 1 that were than compared to the model of each trial to determine whether the trial was responded to correctly in the simulated result.

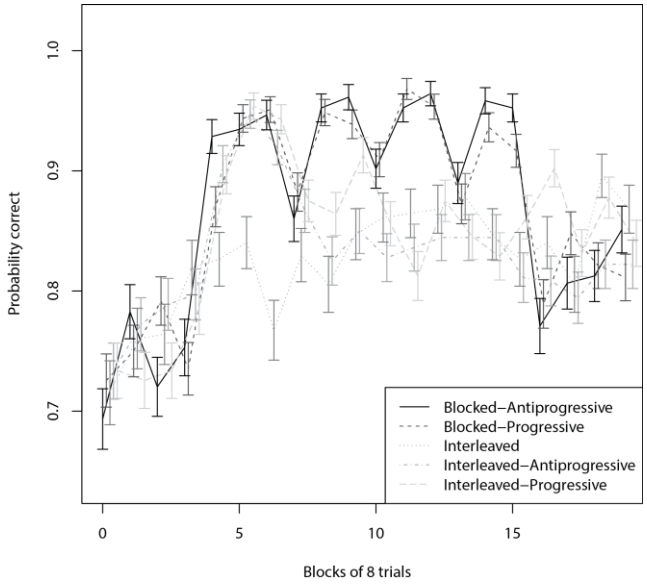

Figure 4. Simulation of experiment

Finally, we wanted to see if there was any general transfer by interval type. While normally we might expect spacing to be an important factor in this effect, the simplicity of experiment (as noted) results in a 50/50 chance of either interval for each trial, so spacing between intervals does not have a great deal of variability. Because of this we used a generalization model that merely tracked the intervals prior practice count, but also used the performance factors analysis (PFA) formalism to track the interval counts depending on success or failure. The PFA method works just like AFM, but counts prior success and prior failure practices instead of simply the count of undifferentiated prior practice [15]. Purely to improve simplicity, we also choose not to account for the fixed interval x tone effects, which did not appear to change the other model coefficients much, despite reducing the fit as expected. Figure 5 shows this model structure. This model adds on 2 PFA components to track learning as a function of prior failures (trif + octf) or success (tris + octs) count for each interval type. Again we fit a single parameter for both intervals under the assumption that are learned at equivalent rates. Again we used the I function to sum the columns since they were mutually exclusive predictors in the equation.

```
answer ~

I(octave0 + octave10 + octave20 + octave30 +
    tritone0 + tritone10 + tritone20 +
tritone30) +

I(soctave0 +
    soctave10 + soctave20 + soctave30 +
stritone0 + stritone10 +
    stritone20 + stritone30) +

I(tris + octs) +

I(trif + octf)
```

Figure 5. GLM model structure with PFA generalization

4.1 Application of Discovered Model to Pedagogical Inference

The model discovered (Figure 5) is useful because it can be used to make pedagogical inference combined with a model of costs for the actions the model allows. The combined model allows us to consider the long-term gains from different conditions of practice relative to the current practice costs. Figure 6 additionally describes a model of practice costs (time spent in practice) as a function of prior practice. This simple model implies a maximal cost for success with unpracticed items, which decreases to a minimum as practice accumulates. This simple model predicts latency cost of success, and we also estimated latency cost of failure and review practice at 7.054 seconds using the overall average from the data. In the model we used the decaying short-term strengths as the predictor.

```
latency ~

I(1/(1 + octave0 + octave10 + octave20 +
    octave30 + tritone0 + tritone10 +
tritone20 + tritone30)
```

Figure 6. LM model structure for costs.

Then we can extract the parameters from both models. See table 3.

Table 3. Parameters used to in pedagogical inference.

Parameter	Value
d	.628
g	.00106
fixed cost failure	7.054
logistic intercept	0.99
spacing coefficient	.131
decay coefficient	2.36
PFA gain failure coeff.	-.106
PFA gain success coeff.	.0144
latency intercept	.154
latency coeff.	1.296

The values from Table 3 then allow us to construct an Excel simulation (available from the first author) of the optimality conditions for our task by examining when spaced gain and PFA

parameter gain (the undecaying learning gains) are maximal relative to the time spent on practice. To do this we plot the learning efficiency (gain / time cost, where both gain and time cost are conditional on success or failure in the calculation) at various levels of prior knowledge (probability values as inferred from the effect of the short-term strength) to find an optimum for the efficiency that allows us to see the optimal probability at which to practice each item. Figure 7 shows the long-term gain curve, the current cost curve, and the optimal efficiency curve.

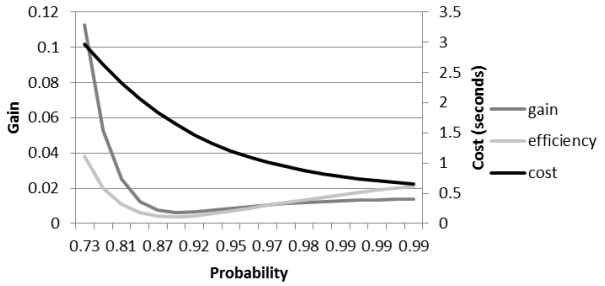

Figure 7. Optimality function.

The results imply it is always best to widely space, since efficiency is maximal with a low probability correct, which would come from very wide spacing between repetitions. However, this may also be because the data was not strong enough to draw conclusions very clearly in this case. Small differences in the fit of the spacing effect result in large changes to the predictions (note the difference in the gain curves for Figures 7 and 8). Indeed, the experiment only had weak power to determine the shape of the spacing effect gain because even in our most intermixed conditions spacing averaged only 8 intervening trials. Because of this the experiment may have poor power when extrapolating to inferences about spacing that imply much wider spacings than were actually used to parameterize the model. As can be seen in Figure 8, when spacing effects are strong the prediction changes.

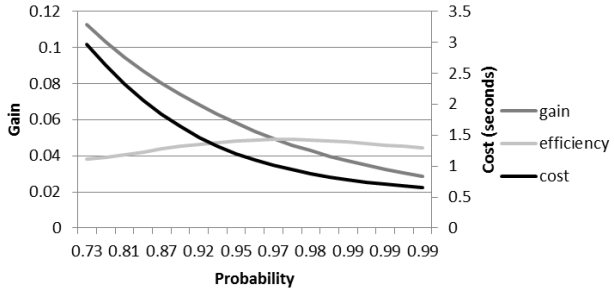

Figure 8. Optimality with spacing parameter (only) changed to .4.

Another problem in trusting the optimality model is due to the noise introduced by only having 2 response options in the experimental design. This makes it hard to identify if success are true success or only guesses, since a guess has a 50% chance of success. This consequently means the PFA parameters are averaging over some of the effects of guessing, thus blunting the quantitative difference between success and failure. This can also have a large effect on the optimization, since the gain curve shape depends on the PFA parameters since they produce the difference in gain for different rates of correctness probability.

So, in addition to what the model reveals about the processes of forgetting and spacing in our data, the model also allows this sort of principled speculation on how the model might be improved if we collect better data to parameterize it. To correct these two limitations of the data we will need to add several more intervals to the practice mix in succeeding experiments. Additional intervals will allow for more spacing (since we have more options for other items to practice when spacing one interval type) and will also improve the effectiveness of the PFA parameters by reducing the noise inherent in modeling success that has a high rate of guessing. In addition to providing more resolution for the spacing effect and PFA parameters, more intervals also will allow much deeper analysis of generalization, since generalization will no longer be a simple binary distinction, but rather a complex categorization.

5. CONCLUSIONS

Although the ability to discern musical intervals is a basic skill vital for almost every musician—beginner or expert—there is a shortage of empirical studies on effective teaching techniques for this skill. In this study, we created a computerized system that tracked participants' identification performance during the process of musical interval learning. The results suggest that our teaching method caused improvement from pretest to posttest, and that an interleaved order was more effective for interval learning. Mathematical models of the data revealed that, while participants improved as a result of our program, there were robust patterns in the practice trials between pretest and posttest. The practice trials showed learning within each block and quick forgetting from one block to the next.

The model we have made is not as detailed as [16], but because of that simplicity, it was possible to more easily fit the model. Importantly, unlike the Pavlik and Anderson model, practices in this new AFM variant model capture spacing effects as permanent learning rather than learning that is merely more durable. Indeed, this new model is in some respects closer to work that has modeled forgetting and spacing using a distribution of units that decay exponentially, some more quickly and some more slowly [14]; however, the model in the current paper is far simpler since it only uses 2 units, one permanent that is a function of current strength and one temporary with relatively quick exponential decay. Others have looked at decay in an educational data mining context, but this work has been at a coarser grain-size looking at forgetting over sessions, and not at the event level [22].

From the perspective of music pedagogy, our training paradigm highlights the importance of the learning sequence in the process of musical interval learning. Ear training and aural skills courses nearly always progress in a rote manner whereby musical relationships are taught serially based on their apparent difficulty (e.g., tones, intervals, chords, harmonies). Our data demonstrate that this typical curriculum is relatively inefficient. Instead, interleaving intervals—here, across multiple pitch registers—seems to promote more efficient learning. Presumably, the higher effectiveness of interleaving in music learning results from having to map sound to meaning across a more diverse acoustic space. Interleaving multiple pitch relationships and registers during the learning processing thus reinforces the learned label for musical sounds across multiple contexts, promoting greater more effective learning. Future work should investigate whether interleaving other musical parameters during an interval learning paradigm (e.g., changes in instrumental timbre) would yield even more robust effects and efficient learning than the changes in register employed presently.

Our paradigm could also be extended to explore novel pitch learning in domains other than music. For example, the effects of spacing and interleaving could be explored in the learning of lexical pitch patterns of tonal languages (e.g., Mandarin Chinese). Unlike English, in these languages, changes in pitch at the syllable level signal word meaning and hence, are entirely novel to nonnative speakers. Future work could thus examine the role of spacing and interleaving in learning the important components of a second language and in maximizing the speed of its acquisition.

6. ACKNOWLEDGMENTS

Our thanks to the University of Memphis for support and funding of this research.

7. REFERENCES

[1] Bidelman, G.M. and Heinz, M.G. 2011. Auditory-nerve responses predict pitch attributes related to musical consonance-dissonance for normal and impaired hearing. *The Journal of the Acoustical Society of America*. 130, 3 (2011), 1488.

[2] Bidelman, G.M. and Krishnan, A. 2009. Neural correlates of consonance, dissonance, and the hierarchy of musical pitch in the human brainstem. *The Journal of Neuroscience*. 29, 42 (2009), 13165–13171.

[3] Draney, K.L. et al. 1995. A measurement model for a complex cognitive skill. *Cognitively diagnostic assessment*. P.D. Nichols et al., eds. 103–125.

[4] Fujioka, T. et al. 2004. Musical training enhances automatic encoding of melodic contour and interval structure. *Journal of cognitive neuroscience*. 16, 6 (2004), 1010–1021.

[5] Hannon, E.E. and Trainor, L.J. 2007. Music acquisition: effects of enculturation and formal training on development. *Trends in Cognitive Sciences*. 11, 11 (2007), 466–472.

[6] Jeffries, T.B. 1967. The Effects of Order of Presentation and Knowledge of Results on the Aural Recognition of Melodic Intervals. *Journal of Research in Music Education*. 15, 3 (1967), 179.

[7] Kameoka, A. and Kuriyagawa, M. 1969. Consonance theory part I: Consonance of dyads. *The Journal of the Acoustical Society of America*. 45, 6 (1969), 1451–1459.

[8] Krumhansl, C.L. 1990. *Cognitive foundations of musical pitch*. Oxford University Press, USA.

[9] Levelt, W.J.M. et al. 2011. Triadic Comparisons of Musical Intervals. *British Journal of Mathematical and Statistical Psychology*. 19, 2 (2011), 163–179.

[10] McDermott, J.H. et al. 2010. Individual differences reveal the basis of consonance. *Current Biology*. 20, 11 (2010), 1035–1041.

[11] McDermott, J.H. et al. 2010. Individual differences reveal the basis of consonance. *Current Biology*. 20, 11 (2010), 1035–1041.

[12] Moreno, S. et al. 2011. Short-term music training enhances verbal intelligence and executive function. *Psychological science*. 22, 11 (2011), 1425–1433.

[13] Moreno, S., and Bidelman, G.M.S. and Research. Understanding neural plasticity Hearing, and cognitive benefit through the unique lens of musical training. *Hearing Research*.

[14] Mozer, M.C. et al. 2009. Predicting the optimal spacing of study: A multiscale context model of memory. *Advances in Neural Information Processing Systems*. Y. Bengio et al., eds. NIPS Foundation. 1321–1329.

[15] Pavlik Jr., P.I. et al. 2009. Performance Factors Analysis -- A New Alternative to Knowledge Tracing. *Proceedings of the 14th International Conference on Artificial Intelligence in Education*. V. Dimitrova and R. Mizoguchi, eds. 531–538.

[16] Pavlik Jr., P.I. and Anderson, J.R. 2005. Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*. 29, 4 (2005), 559–586.

[17] Rubin, D.C. and Wenzel, A.E. 1996. One hundred years of forgetting: A quantitative description of retention. *Psychological Review*. 103, 4 (1996), 734–760.

[18] Samplaski, A. 2005. Interval and interval class similarity: results of a confusion study. *Psychomusicology: Music, Mind & Brain*. 19, 1 (2005), 59–74.

[19] Schellenberg, E.G. 2005. Music and cognitive abilities. *Current Directions in Psychological Science*. 14, 6 (2005), 317–320.

[20] Schellenberg, E.G. and Trainor, L.J. 1996. Sensory consonance and the perceptual similarity of complex-tone harmonic intervals: Tests of adult and infant listeners. *The Journal of the Acoustical Society of America*. 100, (1996), 3321.

[21] Spada, H. and McGraw, B. 1985. The assessment of learning effects with linear logistic test models. *Test design: Developments in psycholgoy and psychometrics*. S. Embretson, ed. Academic Press.

[22] Wang, Y. and Beck, J.E. 2012. Using Student Modeling to Estimate Student Knowledge Retention. *Proceedings of the 5th International Conference on Educational Data Mining*. J. Yacef, K., Zaïane, O., Hershkovitz, H., Yudelson, M., and Stamper, ed. 200–203.