Published in final edited form as:

Atten Percept Psychophys. 2019 January ; 81(1): 344–357. doi:10.3758/s13414-018-1584-x.

Inducing musical-interval learning by combining task practice with periods of stimulus exposure alone

David Little,

Department of Electrical and Computer Engineering, Johns Hopkins University, Barton Hall, 3400 Charles St, Baltimore, MD 21218

Henry Hsu Cheng, and

Department of Communication Sciences and Disorders, Northwestern University, Frances Searle Building, 2240 Campus Dr, Evanston, IL 60208-3550

Beverly Wright

Department of Communication Sciences and Disorders, Knowles Hearing Center, Northwestern Institute for Neuroscience, Northwestern University, Frances Searle Building, 2240 Campus Dr, Evanston, IL 60208-3550

Abstract

A key component of musical proficiency is the ability to discriminate between and identify musical intervals, fixed ratios between pitches. Acquiring these skills requires training, but little is known about how to best arrange the trials within a training session. To address this issue, learning on a musical-interval comparison task was evaluated for two four-day training regimens that employed equal numbers of stimulus presentations per day. A regimen of continuous practice yielded no learning, but a regimen that combined practice and stimulus exposure alone generated clear improvement. Learning in the practice-plus-exposure regimen was due to the combination of the two experiences, because two control groups who received only either the practice or the exposure from that regimen did not learn. Post-test performance suggested that comparison learning generalized to an untrained stimulus type and an untrained musical-interval identification task. Naïve comparison performance, but not learning, was better for larger pitch-ratio differences and for individuals with more musical experience. The reported benefits of the practice-plus-exposure regimen mirror outcomes for fine-grained discrimination and speech tasks, suggesting a general learning principle is involved. In practical terms, it appears that combining practice and stimulus exposure alone is a particularly effective configuration for improving musical-interval perception.

Introduction

A crucial component of musicianship is the ability to discriminate between and identify musical intervals, such as a perfect 4th and a major 3rd. This ability is employed while playing by ear, self-correcting during music performance, and analyzing music (Arenson, 1984; Karpinski, 2000; Cleland & Dobrea-Grindahl, 2013; Furby, 2016). Therefore, learning this skill opens a gateway to musical expertise. Unfortunately, mastery is not easy. Even accomplished performers can struggle to attain proficiency and only people with extensive musical experience tend to reach it (Burns & Ward, 1978; Foxton, Brown, Chambers, &

Griffiths, 2004; McDermott, Keebler, Micheyl, & Oxenham, 2010; Thompson, 2013). Despite the known challenges of musical-interval learning, there has been little investigation of the characteristics of successful training regimens. The focus of the current investigation is on how training influences musical-interval perception depending on the configuration of the training blocks.

Musical-interval learning is so difficult, in part, because of the abstraction required. Musical intervals are determined by the pitch ratio between two musical notes. For example, both the note pair A4 (440 Hz) and D5 (587 Hz) and the pair B2 (123 Hz) and E2 (164 Hz) instantiate a perfect 4th even though their absolute pitches are quite different. Likewise, the pair A4 (440 Hz) and F4 (587 Hz) form a perfect 4th and the pair A4 (440 Hz) and E4 (660 Hz) form a perfect 5th even though they share an absolute pitch. Thus, to identify musical intervals the listener must be able to compute the pitch ratio between two notes well enough to be able to discriminate among different ratios and then apply the appropriate label to each ratio—a challenging categorization task.

We are aware of only two prior training studies related to musical-interval perception. In one (Burns & Ward, 1978), as part of a larger investigation of pitch-ratio discrimination and musical-interval identification, four undergraduate music majors and two musically untrained listeners practiced discriminating between pitch ratios in two-presentation forced-choice trials. Discrimination thresholds were measured separately for each of five different pitch-ratios using an adaptive tracking procedure. Training blocks were repeated until the thresholds reached asymptote. Neither the amount nor temporal distribution of the training was reported. Based on descriptive statistics it appeared that both the music majors and the musically untrained listeners improved and that the music majors reached lower (better) discrimination thresholds. In the other investigation (Foxton et al., 2004), listeners with ~3 or fewer years of musical experience were trained for 25 minutes a day for seven days on melody difference detection. The task was to indicate which of two melodies differed from a random reference melody presented on each trial. During training, for one group (n=10) a single-note change altered the direction of pitch changes—the pitch contour—of the melody. For another group (n=10) a single-note change maintained the pitch contour but changed the pitch ratios of the melody. Between pre- and post-training tests, melody difference detection improved in both groups when the melodies differed in pitch contour, but improved in neither group when the melodies differed in pitch ratio.

Here we compared the effectiveness of two different three-day regimens for training on a musical-interval comparison task that differed only in the configuration of the training blocks: All-Practice and Practice+Exposure. In the All-Practice regimen, listeners practiced the trained task continuously throughout each session, a standard approach to perceptual training. In the Practice+Exposure regimen, in each training session, blocks of task practice were interleaved with blocks of stimulus exposure in the absence of practice. Our motivation was that we had previously observed that Practice+Exposure regimens can yield at least as much learning on perceptual tasks as All-Practice regimens that take the same total time. Thus, with Practice+Exposure training, the amount of practice can be reduced relative to conventional All-Practice training, with no reduction in learning. Practice+Exposure regimens can be successful even when neither the blocks of practice nor the blocks of

stimulus-exposure alone yield learning on their own (as tested with Practice+Silence and Exposure+Silence regimens). In previous investigations, Practice+Exposure training promoted learning on fine-grained discrimination tasks—auditory frequency discrimination (B. A. Wright, Sabin, Zhang, Marrone, & Fitzgerald, 2010) and visual orientation discrimination (Szpiro, Wright, & Carrasco, 2014)—and on speech tasks—acquisition of a non-native phonetic contrast (Wright, Baese-Berk, Marrone, & Bradlow, 2015) and adaptation to a foreign accent (Wright et al., 2015). Therefore, we suspected that it might aid learning on musical-interval perception as well.

During the musical-interval comparison task, two musical intervals were presented and listeners were asked to indicate which of the two was the target interval (a perfect 4^{th}) rather than a foil interval. We chose this task because a listener who cannot distinguish between a target interval and a foil has no hope of labeling those intervals correctly in isolation. Isolated interval identification requires listeners' resolution to be at least a semitone, the smallest difference between two musical intervals, but there are several indications that pitch-ratio discrimination is poorer than that in inexperienced listeners. For example, in one report, thresholds for pitch-ratio discrimination were larger than a semitone and as large as five semitones in listeners with no musical experience, but were smaller than a semitone in formally trained musicians (McDermott et al., 2010). Likewise, during magnitude estimation of pitch-ratios, listeners with less than ~2 years of musical experience often gave the most similar pitch-ratios along a half-semitone grid the same or misordered magnitudes, while listeners with ~11 or more years of experience gave more accurate estimates (Russo & Thompson, 2005). After the training on the musical-interval comparison task was complete, we tested all listeners on the trained comparison condition, the trained comparison task using triangle tones rather than pure tones as during training (an untrained stimulus) and musical-interval identification (an untrained task).

Finally, because all of the listeners completed the same pre-training test, we took advantage of the opportunity to examine naïve performance on musical-interval comparison across a relatively large group of people. Pitch-ratio discrimination performance in naïve listeners has been reported to improve as the difference in pitch-ratio increases, and to be better in musicians than non-musicians (Burns & Campbell, 1994; Burns & Ward, 1978; McDermott et al., 2010). In the present investigation we assessed naïve performance on the related task of musical-interval comparison in a larger group than previously evaluated for discrimination (N=28 compared to N=4–8 per group).

Methods

Listeners

We report data for a total of 28 listeners (21 females) ranging in age from 18 to 27 years (mean 20.2, standard deviation = 2.0). All had normal hearing thresholds at the standard audiometric frequencies as assessed during the investigation (see below), no previous experience with psychoacoustic testing, and no formal musical-interval training. They reported between 0 and 8 years of experience playing a musical instrument (mean = 3.2, SE = 2.8). One listener reported 15 years of experience with voice training. Individuals whose response accuracy on the pre-test (assessed from the first two blocks of trials on the trained

condition on day 1, see below) was ≥ 85% correct (n=9) were excluded from participation because they had relatively little room for improvement (total who completed the pre-test: n=37; total who completed the experiment: n=28). Note that the day 1 performance for one listener was 90% correct (Fig. 2B,C). That listener scored <85% correct on the pre-test (77% correct), but 90% correct across the following 12 blocks of training trials that same day. Listeners were recruited from the Northwestern University community and paid for their participation. All procedures were approved by the Northwestern Office for the Protection of Research Subjects.

General Protocol

All listeners participated in the same basic protocol over four consecutive days. On day 1 there was a familiarization period followed by a pre-test and a training session. On days 2 and 3 the training session was repeated, and on day 4 there was a post-test on the trained condition and two untrained conditions followed by a test of hearing thresholds. The training session differed across four different training regimens, completed by four different listener groups.

Stimuli

The stimuli were tokens of four different musical intervals. Each interval consisted of two sequentially presented pitches that were separated by a constant pitch ratio. The four intervals were a major 3rd ($2^{4/12}$), a perfect 4th (pitch ratio $2^{5/12}$), a perfect 5th ($2^{7/12}$), and a major 6th ($2^{9/12}$), and are referred to as a 3rd, 4th, 5th, and 6th throughout this document. Each pitch pair (interval token) was selected uniformly at random from a fixed set of 24 pitches, from all possible pitch pairs that satisfied the given pitch ratio. These 24 possible pitches fell along the even-tempered scale, starting from the note directly above A4 (440 Hz). In the even-tempered scale, the nth note above A4 has a pitch of 440 X $2^{n/12}$ Hz. Thus pitches ranged from 466 Hz (= 440 × $2^{1/12}$) to 1760 Hz (= 440 × $2^{24/12}$). Because we selected each interval token from a fixed set of 24 pitches, the total number of possible tokens for each interval varied. We chose to select pitches from a fixed range to discourage listeners from using absolute pitch as a cue to differentiate one interval from another.

Each interval token was instantiated by either two pure tones or two triangle tones. The two tones were presented in immediate succession, with the lower pitch first. Each tone was 400 ms in length with 5-ms half-cosine on and off ramps and was presented at 70 dB SPL. All stimuli were generated using a Roland Cakewalk US-25EX audio interface at a sampling rate of 44,100 Hz and a bit rate of 16, and delivered to the left ear over Sennheiser HD265 headphones with circumaural cushions. Listeners were tested in a sound attenuated booth.

Trained Task

The trained task was musical-interval comparison. On each two-presentation forced-choice trial, listeners were presented a target musical interval (a 4th) and one of three foil intervals (a 3rd, 5th or 6th). The target and foil intervals were presented in random order and were marked on a computer screen as “Interval 1” and “Interval 2.” The two presentations were separated by 900 ms, measured from the offset of the second tone in the first presentation to the onset of the first tone in the second presentation. Listeners were asked to indicate which

of the two presentations contained the 4th by pressing a key on a computer keyboard. They were allowed to respond at their own pace. Feedback displayed on the computer screen after each response indicated whether the response was 'Correct!' or 'Wrong!' and provided the appropriate labels (3rd, 4th, 5th or 6th) for the intervals that had been presented in each observation period. The feedback remained on the screen for 1350 ms and then the next trial began. The mean trial length, including listener response time, was ~4300 ms. At the beginning of each block of trials, listeners were given two randomly selected tokens of a 4th, each labeled as a 4th.

Training Regimens

Each of the four training regimens consisted of 12 30-trial training blocks per day for three consecutive days, for a grand total of 360 × 3 = 1080 training trials. They were composed of various combinations of three block types: practice, exposure and silence. During the *practice* blocks (Fig. 1 – left panel), listeners performed the musical-interval comparison task with pure-tone stimuli. During the *exposure* blocks (Fig. 1 – middle panel), listeners were presented randomly generated tokens of pure-tone 4ths in the background while they performed a written symbol-to-number matching task. Stimulus presentations were organized into "trials" modeled after the practice trials, but in which only 4ths were presented. The purpose of the written matching task was to engage the listeners in a non-auditory task while the stimulus exposures were presented. For that task, listeners were provided with a sheet of paper at the top of which was a table of 80 randomly selected wingding font symbols each associated with a number from 1 to 80. At the bottom of the sheet was a random sequence of 160 instances of those 80 symbols. Listeners were asked to write the corresponding number next to each symbol sequentially from the top to the bottom of the sheet and to continue to another sheet if they finished before the interval presentations were complete. During the *silence* blocks (Fig. 1 – right panel), listeners performed the same task as for the practice blocks (musical-interval comparison), except that no interval tokens were presented. Listeners were told that they would hear no sounds, and were asked to simply do their best to "guess" what the randomly generated correct answer would be. They received the same feedback as for the practice blocks. Thus, the experience during the silence blocks was as similar as possible to the practice blocks, but without any sound stimuli.

The four training regimens provided different amounts of practice and exposure (Figure 3A – right column). To document the effect of continuous practice, in the *All-Practice* regimen, all 12 blocks in each session were practice blocks (360 practice trials per session; n=7). To determine the effect of replacing half of the practice trials with exposure alone, in the *Practice+Exposure* regimen the 12 blocks in each session alternated every 2 blocks between practice and exposure blocks, for a total of 6 blocks each, starting with practice (180 practice trials+180 exposure trials per session; n=7). To establish the influence of the practice portions of the Practice+Exposure regimen, in the *Practice+Silence* regimen the 12 blocks in each session alternated every 2 blocks between practice and silence blocks, for a total of 6 blocks each, starting with practice (180 practice trials+180 silence trials per session; n=7). Likewise, to ascertain the influence of the exposure-alone portions of the Practice+Exposure regimen, in the *Exposure+Silence* regimen the 12 blocks in each session alternated every 2

blocks between exposure and silence blocks, for a total of 6 blocks each, starting with silence (180 exposure trials+180 silence trials per session; n=7).

Listeners were assigned to groups as follows. The first four listeners completed the Practice +Exposure regimen and the next three completed the Practice+Silence regimen. Subsequent listeners were assigned at random to the four regimens. There was no significant relationship between percent correct on day 4 and the date of the first day of training either before ($p = 0.384$) or after ($p = 0.657$) controlling for any influence of percent correct on day 1, according to a single-level logistic regression (following the same approach described in the analysis section below).

Familiarization

On day 1, listeners were first familiarized with the concept of a musical interval. To do so, they were shown schematic spectrograms of single tokens of each of the four musical intervals, each with a different starting pitch, and told that the relationship between pitches, not the absolute pitch, was the important feature of the stimulus. They were then given a verbal description of the interval-comparison task, followed by four presentations of randomly selected tokens of a 4^{th}, each labeled as a 4^{th}, 8 practice trials with feedback, a second set of four tokens of a 4^{th}, and a second set of 8 practice trials.

Pre- and post-tests

The pre- and post-tests were the same for all four regimens. The pre-test occurred on day 1 immediately after the familiarization period and consisted of two 30-trial practice blocks with feedback. These blocks were in addition to the 12 daily training blocks that occurred on the same day.

The post-test occurred on day 4 and consisted of four 30-trial blocks of each of three conditions: musical-interval comparison with pure tones (trained task and stimulus), musical-interval comparison with triangle tones (untrained stimulus), and musical-interval identification with pure tones (untrained task). During *interval comparison with triangle tones*, trials proceeded as in the trained task, except that each pitch ratio was composed of triangle tones, rather than pure tones. A triangle tone is defined as $|(4ft) \bmod 4 - 2| - 1$, where f is the frequency in Hz and t is the time in seconds. During *interval identification*, on each trial, listeners were presented with a single pure-tone interval token and asked to identify which interval had been presented: a 3^{rd}, 4^{th}, 5^{th}, or 6^{th}. Each token was randomly selected from the four possible pitch ratios with equal probability. The four options for the interval label were displayed on a computer screen and listeners were asked to indicate which interval they heard by pressing a key on a keyboard. Following each response, feedback displayed on the computer screen indicated whether the response was ‘Correct!’ or ‘Wrong!’ and indicated which of the four intervals had been presented. The feedback remained on screen for 1350 ms before the next trial began. Listeners completed the interval comparison trials with pure tones first, and then completed the other two conditions in counterbalanced order. Finally, we verified that listeners had normal pure-tone thresholds at the standard audiometric frequencies as assessed with a two-interval forced-choice procedure.

Analysis

We analyzed the data using single-level regressions for the mean data and a multiple-level (mixed-effects) regression for the individual data. For the mean data, the dependent variable was the proportion of correct responses (ranging from 0 to 1, percent correct = proportion correct x 100) and for the individual data it was the response accuracy on a given trial (either 0 or 1). The regressions employed a logistic curve, which can be used to model bounded quantities (between 0 and 1) or discrete outcomes (0 or 1). For the single-level regressions, the mean proportion correct was computed for each listener and level of the within-listener predictors (e.g., day). These means were then fitted by each regression using all predictors, both the within-listener (e.g., day) and across-listener (e.g., regimen) ones. Which specific within- and across-listener predictors were included depended on the specific analysis. All single-level regressions of the data from the trained condition were fitted by regimen and day. The regressions of the data from the untrained post-test conditions were fitted either by the regimen or by the trained-condition post-test. Additional predictors included for specific analyses are noted in the results (for example, a more detailed analysis of the training data included foil as a predictor). For the multiple-level regression, the individual data were modeled with two levels: one level modeled the effects of each individual listener and the other the effects of the regimen. Each level contributed to the slope and intercept of a line fitted across day for each comparison foil. Note that any group differences revealed by this analysis were present in spite of any individual variation captured by the model, such as that due to starting performance.

To implement these regressions we used a Bayesian approach for its flexibility (Gelman & Hill, 2007) and for its conservative nature when N is small. In a Bayesian analysis, a probability distribution, called the posterior is computed for each parameter of the model (e.g., the mean of the distribution and the variance of the distribution). The posterior is based on a prior, which is an assumed distribution of the parameter, and on the data themselves. The prior is selected to shrink the posterior towards zero. Thus, when compared to a more traditional regression analysis, a more conservative estimate is computed when N is small. To compute the posterior, we used the Stan modeling language (version 2.7.0) to find 4000 MCMC samples (Hoffman & Gelman, 2014) (4 chains, of 2,000 samples each with the first 1,000 burn-in samples of each chain discarded). This number of samples was selected to keep the estimated error of the method acceptably low (≤ 0.28 percentage points of the dependent variable). These posterior samples were then used to compute the p-values, means and standard errors reported in the results. In the few cases in which the standard error was a nonsensical quantity, because the posterior distribution was asymmetric, the highest posterior-density interval with an equivalent density to one standard error (~68.2%) is shown in the figures. Where appropriate, p-values were corrected for multiple comparisons (Gelman & Tuerlinckx, 2000; Gelman, Hill, & Yajima, 2012).

For the single-level regressions, to ensure stability and robustness, the dependent variable (y) was transformed by r/2 + y(1-r)—where r was a relatively small proportion of the data range (0 to 1) modeled to best fit the data. For further robustness, for both the single- and multi-level regressions, we used a beta distribution to model overdispersion. The prior over r was a zero-mean normal distribution (truncated below 0) with a variance of 5 percentage points,

and the prior over the beta was an uninformative zero-mean normal distribution (truncated below 0) with a variance of 100 (which is effectively uniform given the range of the data). Priors over the coefficients of the logistic regression were zero-mean Cauchy distributions with a variance of 5 (Gelman, Jakulin, Pittau, & Su, 2008). Hyperpriors in the multilevel regression were weakly informative Cauchy distributions (scale 1) (Gelman, 2006). We validated the model designs and choice of priors using posterior predictive checks (Gelman, Meng, & Stern, 1996). Those checks indicated that the resulting models accurately predicted the 97.5%, 84.1%, 50%, 15.9% 2.5% percentiles of the residuals with a standard error of ≤ 2.8 percentage points (posterior predictive $p \geq 0.248$).

Code Availability

The full analyses and all experimental data are available at https://github.com/haberdashPI/apmusic.

Results

Pre-training performance

Initial response accuracy on the interval-comparison task differed across the foils (4th vs. 3rd, 5th or 6th) and improved with increasing musical experience. Figure 2A shows the percentage of correct responses for each of the 28 listeners (raw data; gray circles) for the three foils (x axis) and the means and standard errors for these data estimated from a multi-level regression across all days of training (bars and error bars; see Methods). Comparison performance was greater for the 4th versus the 6th (73.9%, SE = 5.3) than for the 4th versus the 3rd (63.6%, SE = 3.8) and the 4th versus the 5th (61.2%, SE = 4.5) (multi-level regression: $p \leq 0.041$), and did not differ between the latter two cases ($p = 0.510$). Figure 2B shows individual starting performance (raw data: grey circles) for the three pairwise comparisons (x- and y-axes) among the foils (panels), and the means and standard errors from the multi-level regression (lines and gray regions). Starting performance was significantly correlated between each pair of foils, such that the regression coefficients for one foil could account for ≥ 53% of the variance in the coefficients for another foil (multi-level regression: $r \geq 0.73$, SE ≤ 0.12, $p \leq 0.001$). Figure 2C shows the number of years of musical experience (x-axis) and percent-correct performance (y-axis) for each listener (raw data; symbols), and the fit and standard error across years of musical experience according to the multi-level regression (line and gray region). Each year of experience improved the odds in favor of a correct response on day 1 by 10% (SE = 3.1, $p < 0.005$). Note that, given the odds (o) the probability (p) of a correct response is $p = o / (1+o)$.

Trained Condition

Only the Practice+Exposure regimen yielded learning on musical-interval comparison at the group level. Figure 3A shows the mean percent-correct performance for each group (raw data; symbols with standard error) across the four days of training and the fit and standard error of a single-level regression of day and group (lines and shaded regions). Accuracy for the Practice+Exposure group (black square) improved from 68.9% (SE = 4.1) on day 1 to 88.2% (SE = 2.7) on day 4 (single-level regression: $p = 0.002$), but did not improve for any of the other groups: All-Practice (gray square), Practice+Silence (gray circle), or Exposure

+Silence (white circle) ($p \geq 0.142$, starting at 61.7–65.7%, SE ≤ 5.1 and ending at 71.5–71.9%, SE ≤ 4.8). Day 1 responses did not differ across the groups (absolute difference ≤ 6.6 percentage points, SE ≤ 6.1, $p \geq 0.272$). By the end of training, the performance of the Practice+Exposure group was ≥ 16.3 percentage points higher than that of the remaining three groups (SE ≤ 5.5, $p \leq 0.003$), whose performance did not differ from one another (absolute difference ≤ 0.7 percentage points, SE ≤ 6.3, $p \geq 0.927$). The amount of learning did not differ across the three foils (4th vs. 3rd, 5th or 6th) for any of the groups (single-level regression of day, group and foil: absolute difference ≤ 2.2 percentage points, SE ≤ 7.8, $p \geq 0.786$), despite the differences in starting performance across foil (see above).

The advantage of the Practice+Exposure group over the other three groups remained even after accounting for individual differences in starting performance (multi-level regression: $p \leq 0.046$). At the individual level, more listeners learned, and tended to learn more, in the Practice+Exposure group than in the other three groups. Figure 3B shows the performance across days for each individual listener (raw data: symbols and dotted lines), separated by training group (panels), and the fit and standard error of the multi-level regression to these data (lines and shaded regions). Listeners who improved significantly from day 1 to day 4 (by the multi-level regression) were termed learners (black lines) and the others non-learners (dark gray lines). All 7 listeners in the Practice+Exposure group learned (multi-level regression: $p \leq 0.001$). In contrast, only about half of the 7 listeners improved in each of the other groups (n=3 or 4; $p \leq 0.013$). Further, among the listeners who improved, the amount of improvement tended to be larger in the Practice+Exposure group (15.7–23.3 percentage points, SE ≤ 4.1, as estimated from the regression) than in the other groups: All-Practice (n=3; 11.2–20.4, SE ≤ 4.4), Practice+Silence (n=3; 9.9–14.0, SE ≤ 4.0), and Exposure +Silence (n=4; 11.7–16.4, SE ≤ 5.5). Listeners identified as non-learners improved by less than 3.1 percentage points (SE ≤ 4.2, $p \geq 0.164$). The amount of musical experience did not appear to affect the results. Musical experience did not differ across the groups (ANOVA of experience across group: $F(3,24) = 0.1$, $p = 0.969$, non-parametric Kruskal-Wallis $X2(3) = 0.6$, $p= 0.891$). It did not predict the amount of listener improvement in the multi-level regression ($p \geq 0.721$) and its inclusion in that regression did not influence the overall outcome.

Finally, Figure 3C shows the performance of individual listeners (raw data: symbols) on day 1 (x axis) and day 4 (y axis), and the fits by single-level regression for each group (lines). Consistent with the outcome of the multi-level analysis, the data points for the majority of listeners in the All-Practice (gray diamonds), Practice+Silence (dark gray triangles), and Exposure+Silence (white circles) groups fell close to the diagonal (no learning) or below it (worsening; light gray region). In contrast, the data points for all of the listeners in the Practice+Exposure group (black squares) were well above the diagonal (learning), regardless of starting performance.

Untrained conditions

Untrained stimulus: Interval comparison with triangle tones—Post-test performance on musical-interval comparison with an untrained stimulus (triangle tones) essentially matched that with the trained stimulus (pure tones). Figure 4A shows the percent-

correct performance for each individual listener (raw data: gray circles), separated by training group, and the means and standard errors determined by a single-level regression across stimulus and group (bars and error bars) for the pure-tone (solid bars) and triangle-tone (striped bars) stimuli. Within each group, the response accuracies did not differ between the two stimulus types (single-level regression: absolute difference ≤ 2.2 percentage points, SE ≤ 5.9, $p \geq 0.711$). Therefore, the Practice+Exposure group outperformed the other three groups with the untrained triangle tones (≥ 15.4, SE ≤ 5.0, $p \leq 0.002$), just as with the trained pure tones (≥ 16.3, SE ≤ 4.9, $p \leq 0.002$). Figure 4B shows percent correct performance for the pure tones (x-axis) and the triangle tones (y-axis) for each listener in each group (raw data; symbols), and the fit and standard error of a single-level regression of triangle-tone by pure-tone response accuracy (line and shaded area). Performance with the pure tones accounted for 80.3% of the variance in performance with the triangle tones (single-level regression: SE = 1.7 $p < 0.001$). The relationship between the response accuracy with pure tones and triangle tones was not influenced by the group or foil (group: single-level regression of triangle-tone responses across pure-tone responses and group: absolute difference ≤ 2.8 percentage points, SE ≤ 4.9, $p \geq 0.536$; foil: single-level regression of triangle-tone responses across pure-tone responses and foil: absolute difference ≤ 4.6 percentage points, SE ≤ 10.7, $p \geq 0.617$). Thus, regardless of group or foil, performance with the untrained triangle-tone stimulus was well predicted by performance with the trained pure-tone stimulus.

Untrained task: Interval identification—Post-test performance on the untrained interval-identification task was better for the Practice+Exposure group than for the other three groups. Figure 5A shows the percentage of correct responses on the identification task for each listener (gray circles), separated by training group, and the means and standard errors determined by a single-level regression across group (bars and error bars). The overall identification accuracy for the Practice+Exposure group was 55.9% (SE = 3.9), ≥ 14.3 percentage points higher than for each of the other three groups (single-level regression: SE = 5.5, $p \leq 0.015$), among which performance did not differ (absolute difference ≤ 7.6 percentage points, SE ≥ 5.4, $p \geq 0.183$). Figure 5B shows the post-test response accuracy (raw data; shades and symbols) for the trained interval-comparison task (x-axis) and untrained identification task (y-axis) and the fit and standard error of a single-level regression of identification by comparison response accuracy (line and shaded region). Comparison performance at the post-test accounted for 52.7% of the variance in identification performance (single-level regression: SE = 3.9, $p < 0.001$; without individual at 87.6% correct comparison performance: $R^2 = 0.508$, SE = 0.039, $p < 0.001$). The relationship between identification and comparison accuracy did not differ across the groups (single-level regression of identification accuracy across comparison accuracy and group: absolute difference ≤ 10.1 percentage points, SE ≤ 8.6, $p \geq 0.249$). Figure 5C is plotted as for Figure 5B but with post-test response accuracy shown separately for each interval (raw data; symbols) and with the fit and standard error of a single-level regression of identification response accuracy by comparison response accuracy and interval (lines and shaded regions). For comparison response accuracies above 64.5%, identification performance was better for the perfect 4th than for the other three intervals (single-level regression: $p \leq 0.05$). Therefore, it appears that regardless of group, performance at the end

of training on the trained comparison task determined how well listeners performed on the untrained identification task, especially when identifying the interval that was the target during comparison training (the 4th).

Discussion

The purpose of the present study was to compare the amount of learning on a musical-interval comparison task generated by two different training regimens. The task was to indicate which of two musical intervals was the target interval (a perfect 4th), rather than one of three foils (a major 3rd, perfect 5th and major 6th). An All-Practice group who performed the task continuously throughout each of three consecutive daily practice sessions did not learn. In contrast, a Practice+Exposure group who performed the task only intermittently during each session, but who received stimulus exposure alone during the remaining periods, did learn. This learning resulted from the combination of practice and exposure because two control groups did not learn: a Practice+Silence group who only performed the task intermittently, with no periods of stimulus-exposure alone, and an Exposure+Silence group who only received the intermittent periods of stimulus-exposure alone, but did not perform the task. After training, the performance of the Practice+Exposure group was also better than that of the other groups on musical-interval comparison with an untrained triangle-tone stimulus and on an untrained musical-interval identification task. Finally, the magnitude of improvement was similar for all three foils (3rd, 5th and 6th)—despite differences in comparison performance across them on the first day—and had no clear relationship with the number of years of musical experience—despite the advantage more years of experience provided on the first day. Overall, the results suggest that the combination of task practice and stimulus-exposure alone is a promising means to improve musical-interval perception and more generally that the configuration of training blocks is crucial to the outcome of musical-interval training.

Effects of practice and stimulus exposure alone

The present results extend the demonstrations of the effectiveness of combining task practice and stimulus exposure alone from fine-grained discrimination (Szpiro et al., 2014; B. A. Wright et al., 2010) and speech tasks (Wright et al., 2015) to a challenging nonspeech classification task. That this combination training aids learning on such a wide range of tasks suggests that the mechanisms engaged by it are quite general. One proposal is that there are at least two requirements for long-lasting perceptual learning: a requirement for sufficient sensory stimulation—provided by stimulus exposures with or without associated task practice—and another requirement for an internal permissive signal such as attention or reward—provided by task practice but not stimulus exposure alone (Wright et al., 2015; B. A. Wright et al., 2010). The idea is that stimulus exposure alone cannot generate long-lasting perceptual learning on its own because such exposures lack this permissive signal (Ahissar & Hochstein, 2004; Gilbert & Sigman, 2007; Aaron R Seitz & Dinse, 2007; A. Seitz & Watanabe, 2005; Wright & Sabin, 2007; Wright & Zhang, 2009); however, the combination of task practice and stimulus exposure alone can yield perceptual learning because (1) the periods of task practice lead to effects that remain during subsequent periods of stimulus exposure alone, allowing the exposures to contribute to learning and/or (2) the

periods of exposure, though unable to contribute to learning directly, increase the potential for subsequent periods of task practice to generate learning (B. A. Wright et al., 2010).

The current data also demonstrate that the combination of task practice and stimulus exposure alone can exceed the benefits of an equivalent period of task practice, as has been reported previously (Szpiro et al., 2014; Wright et al., 2015). One potential explanation for this finding is that practice—but not stimulus exposure alone—slowly depletes a limited resource that is necessary for learning and that is replenished during practice breaks. Similar explanations have been proposed to account for the benefits of distributed over massed training (Naqib, Sossin, & Farah, 2012; Pavlik & Anderson, 2005). New here is the idea that stimulus exposure alone neither draws on that limited resource nor prevents that resource from recovering, even though this exposure can contribute to learning when paired with task practice. Note that this proposed limited resource differs from the proposed permissive signal. The permissive signal is triggered by task practice and its effects can extend to periods of stimulus exposure alone, while the limited resource is depleted by task practice and is not influenced by mere stimulus exposure.

While the key outcome reported here is the benefit of combining task practice with stimulus exposure alone, at least four aspects of the practice periods themselves may have contributed to the learning observed. First, the task was musical-interval comparison rather than single-interval identification as in traditional musical aural training (e.g. Karpinski, 2000). The direct comparison between stimuli that this task required may have helped emphasize the differences between similar pitch-ratios and thereby aided learning. Second, the feedback after each trial indicated not only whether the response was correct or incorrect—as is typical for a discrimination task—but also provided the names of the target and foil intervals. This joint feedback may have facilitated the formation of musical-interval categories, potentially by targeting a rule-based learning system for which such feedback can facilitate category learning (Ashby & Maddox, 2005, Yi & Chandrasekaran 2016). Third, on a given trial, compared to the pitch ratio of the target interval (a 4^{th}), the pitch ratio of the foil interval could be either smaller (a 4^{th} vs. a 3^{rd}) or larger (a 4^{th} vs. a 5^{th} or 6^{th}), rather than consistently larger as in the tasks typically used to assess pitch-ratio discrimination (Burns & Campbell, 1994; Burns & Ward, 1978; McDermott et al., 2010). The task thus required reference to an absolute pitch ratio, rather than simply the selection of the larger of two pitch ratios, which may have supported learning of the absolute pitch-ratio of the target. Fourth, the pitch ratio of the target interval remained constant throughout the experiment, rather than varying as in traditional musical aural training (Karpinski, 2000). It is possible that training multiple different targets in a single session would have impeded learning on any one target, as has been documented for a number of perceptual discrimination tasks (Aberg & Herzog, 2009; Banai, Ortiz, Oppenheimer, & Wright, 2010; Maidment, Kang, Gill, & Amitay, 2015; Parkosadze, Otto, Malania, Kezeli, & Herzog, 2008; Aaron R. Seitz et al., 2005; Yotsumoto, Chang, Watanabe, & Sasaki, 2009).

Just as some specific aspects of the practice periods may have contributed to the present results, the particular nature of the stimulus exposure alone may also have been important. Listeners were presented two tokens of a 4th on each trial during stimulus exposure alone, but were presented only a single token of a 4^{th} paired with a single token of a foil (3rd, 5th

or 6th) on each trial during practice. Thus, during stimulus exposure alone there were twice as many presentations of the target interval (the 4th) than during practice, and no presentations of the foil intervals. These differences may have contributed to the greater effectiveness of the Practice+Exposure compared to the All-Practice regimen. Even if that is the case, however, their influence was not sufficient for the Exposure+Silence and Practice +Silence regimens to yield different outcomes.

Generalization across stimulus and task—The similarity in the mean post-test comparison performance between the trained (pure tone) and untrained (triangle tone) stimulus, as well as the strong correlation in individual performance at the post-test between those two stimulus types, suggests that learning generalized across stimulus type. The implication is that training modified neural circuitry that was responsive to the pitch-ratio of the stimulus and at least somewhat insensitive to the stimulus type. Cross-stimulus generalization has been documented on other auditory discrimination tasks, though it is relatively rare following training on a single condition (for reviews see Fahle & Poggio, 2002; Wright & Zhang, 2009).

Learning also appeared to have generalized from interval comparison to interval identification, because the group that learned the most during comparison training (Practice +Exposure) performed the best on average on the identification task during the post-test, and because performance on the comparison and identification tasks was strongly correlated across individuals. If so, the generalization was greatest for the pitch-ratio of the trained musical interval given the better identification performance for the target interval than the foil intervals. The implication here is that training modified neural circuity contributing to both the comparison and identification of musical intervals. Others have reached a similar conclusion concerning the relationship between discrimination and identification based on evidence that the discrimination of pitch-ratios is better across rather than within musical-interval category boundaries (Burns & Campbell, 1994; Burns & Ward, 1978). A shared process for discrimination and identification of speech sounds has also been implicated because discrimination training and identification training both led to similar improvements on discrimination (Flege, 1995; Wayland & Li, 2008), and discrimination training led to improvements on various speech classification-related tasks (Fu, Galvin, Wang, & Nogaki, 2005; Moore, Rosenberg, & Coleman, 2005; e.g. Pisoni, Aslin, Perey, & Hennessy, 1982; Strange & Dittmann, 1984).

Finally, the strong correlation in performance across all listeners at the post-test between the trained and untrained stimulus type, and between the trained and untrained tasks, suggests that the important factor for generalization was the amount of learning on the trained condition irrespective of which training regimen induced that learning. It therefore appears that the different training regimens may have modified the same locus of learning, but to varying degrees.

Naive interval-comparison performance

Before training, the present listeners were better able to identify which of two intervals was the target interval when the comparison was to foils with the largest pitch-ratio difference

from the target (4^{th} vs. 6^{th}) as opposed to foils with pitch-ratios that were more similar to the target (4^{th} vs. 3^{rd} and 4^{th} vs 5^{th}). This outcome is consistent with prior studies of pitch-ratio discrimination performance in listeners with varying degrees of experience (Burns & Ward, 1978; Burns & Campbell, 1994; McDermott et al., 2010). Further, the performance of the present listeners was highly correlated between any given pair of foil intervals. At least some of the across-listener variation can be attributed to differences in musical experience: before training, musicians were better at the musical-interval comparison task than non-musicians, as has been reported previously for pitchratio discrimination (Burns & Ward, 1978; Burns & Campbell, 1994; McDermott et al., 2010). This musician advantage may actually be underestimated in the present case, because the inclusion criteria (≤ 85% pre-test accuracy and no formal interval-training experience) likely excluded more musicians than non-musicians. Thus, it appears that musical intervals are ordered perceptually by their pitch-ratio magnitude and that before training the capacity to recognize specific intervals in direct comparison differs across listeners due to systematic variation in an ability that is enhanced by musical experience.

Summary

1) Combining periods of task practice and stimulus exposure alone led to better learning on musical-interval comparison than continuous practice, demonstrating that the configuration of training trials can influence the amount of musical-interval learning.

2) Training on musical-interval comparison led to better identification of those same intervals, suggesting that difficulties in identification might be mitigated with interval-comparison training.

3) Neither initial performance nor musical experience affected the benefits of combining practice on musical-interval comparison and stimulus exposure alone, suggesting that such training may be useful to listeners at a variety of experience levels.

Acknowledgments

This work was supported in part by the National Institute on Deafness and Other Communication Disorders–National Institutes of Health, the Northwestern University Bioscientist program funded by HHMI Undergraduate Science Education Grants to Research Universities #52006934, and DARPA. We thank Jessica Conderman for providing valuable input in the design of the trained task.

References

Aberg KC, & Herzog MH (2009). Interleaving bisection stimuli – randomly or in sequence – does not disrupt perceptual learning, it just makes it more difficult. Vision Research, 49(21), 2591–2598. [PubMed: 19616572]

Ahissar M, & Hochstein S (2004). The reverse hierarchy theory of visual perceptual learning. Trends in Cognitive Sciences, 8(10), 457–464. [PubMed: 15450510]

Arenson MA (1984). Computer-based instruction in musicianship training: Some issues and answers. Computers and the Humanities, 18(3–4), 157–163.

Ashby FG, & Maddox WT (2005). Human Category Learning. Annual Review of Psychology, 56(1), 149–178.

Banai K, Ortiz JA, Oppenheimer JD, & Wright BA (2010). Learning two things at once: Differential constraints on the acquisition and consolidation of perceptual learning. Neuroscience, 165(2), 436–444. [PubMed: 19883735]

Burns EM, & Campbell SL (1994). Frequency and frequency-ratio resolution by possessors of absolute and relative pitch: Examples of categorical perception? The Journal of the Acoustical Society of America, 96(5), 2704–2719. [PubMed: 7983276]

Burns EM, & Ward WD (1978). Categorical perception—phenomenon or epiphenomenon: Evidence from experiments in the perception of melodic musical intervals. The Journal of the Acoustical Society of America, 63(2), 456–468. [PubMed: 670543]

Cleland KD, & Dobrea-Grindahl M (2013). Developing Musicianship Through Aural Skills: A Holistic Approach to Sight Singing and Ear Training. Routledge.

Fahle M, & Poggio T (2002). Perceptual learning. MIT Press.

Flege JE (1995). Two procedures for training a novel second language phonetic contrast. Applied Psycholinguistics, 16, 425–442.

Foxton JM, Brown ACB, Chambers S, & Griffiths TD (2004). Training improves acoustic pattern perception. *Current* Biology: CB, 14(4), 322–325. [PubMed: 14972683]

Fu Q-J, Galvin J, Wang X, & Nogaki G (2005). Moderate auditory training can improve speech performance of adult cochlear implant patients. Acoustics Research Letters Online, 6(3), 106–111.

Furby VJ (2016). The Effects of Peer Tutoring on the Aural Skills Performance of Undergraduate Music Majors. Update: Applications of Research in Music Education, 34(3), 33–39.

Gelman A (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Analysis, 1(3), 515–534.

Gelman A, & Hill J (2007). *Data analysis using regression and* multilevel/hierarchical models (Vol. 1). Cambridge University Press New York, NY, USA.

Gelman A, & Tuerlinckx F (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. Computational Statistics, 15(3), 373–390.

Gelman A, Hill J, & Yajima M (2012). Why We (Usually) Don't Have to Worry About Multiple Comparisons. Journal of Research on Educational Effectiveness, 5(2), 189–211.

Gelman A, Jakulin A, Pittau MG, & Su Y-S (2008). A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models. The Annals of Applied Statistics, 2(4), 1360–1383.

Gelman A, Meng X-L, & Stern H (1996). Posterior predictive assessment of model fitness via realized discrepancies. Statistica Sinica, 6(4), 733–760.

Gilbert CD, & Sigman M (2007). Brain States: Top-Down Influences in Sensory Processing. Neuron, 54(5), 677–696. [PubMed: 17553419]

Hoffman MD, & Gelman A (2014). The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. Journal of Machine Learning Research, 15(Apr), 1593–1623.

Karpinski GS (2000). Aural skills acquisition: The development of listening, reading, and performing skills in college-level musicians. Oxford University Press.

Maidment DW, Kang H, Gill EC, & Amitay S (2015). Acquisition versus Consolidation of Auditory Perceptual Learning Using Mixed-Training Regimens. PLoS ONE, 10(3), e0121953. [PubMed: 25803429]

McDermott JH, Keebler MV, Micheyl C, & Oxenham AJ (2010). Musical intervals and relative pitch: Frequency resolution, not interval resolution, is special. The Journal of the Acoustical Society of America, 128(4), 1943–1951. [PubMed: 20968366]

Moore DR, Rosenberg JF, & Coleman JS (2005). Discrimination training of phonemic contrasts enhances phonological processing in mainstream school children. Brain and Language, 94(1), 72–85. [PubMed: 15896385]

Naqib F, Sossin WS, & Farah CA (2012). Molecular Determinants of the Spacing Effect. Neural Plasticity, 2012, 1–8.

Parkosadze K, Otto TU, Malania M, Kezeli A, & Herzog MH (2008). Perceptual learning of bisection stimuli under roving: Slow and largely specific. Journal of Vision, 8(1), 5.

Pavlik PI, & Anderson JR (2005). Practice and Forgetting Effects on Vocabulary Memory: An Activation-Based Model of the Spacing Effect. Cognitive Science, 29(4), 559–586. [PubMed: 21702785]

Pisoni DB, Aslin RN, Perey AJ, & Hennessy BL (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. Journal of Experimental Psychology: Human Perception and Performance, 8(2), 297–314. [PubMed: 6461723]

Russo FA, & Thompson WF (2005). The subjective size of melodic intervals over a two-octave range. Psychonomic Bulletin & Review, 12(6), 1068–1075. [PubMed: 16615330]

Seitz AR, & Dinse HR (2007). A common framework for perceptual learning. Current Opinion in Neurobiology, 17(2), 148–153. [PubMed: 17317151]

Seitz AR, Yamagishi N, Werner B, Goda N, Kawato M, & Watanabe T (2005). Task-specific disruption of perceptual learning. Proceedings of the National Academy of Sciences of the United States of America, 102(41), 14895–14900. [PubMed: 16203984]

Seitz A, & Watanabe T (2005). A unified model for perceptual learning. Trends in Cognitive Sciences, 9(7), 329–334. [PubMed: 15955722]

Strange W, & Dittmann S (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. Perception & Ps*ychophysics*, 36(2), 131–145. [PubMed: 6514522]

Szpiro SFA, Wright BA, & Carrasco M (2014). Learning one task by interleaving practice with another task. Vision Research, 101, 118–124. [PubMed: 24959653]

Thompson WF (2013). 4 - Intervals and Scales In Deutsch D (Ed.), The Psychology of Music *(*Third *Edition)* (pp. 107–140). Academic Press.

Wayland RP, & Li B (2008). Effects of two training procedures in cross-language perception of tones. Journal of Phonetics, 36(2), 250–267.

Wright BA, & Sabin AT (2007). Perceptual learning: How much daily training is enough? Experimental Brain Research, 180(4), 727–736. [PubMed: 17333009]

Wright BA, & Zhang Y (2009). A Review of the Generalization of Auditory Learning. Philosophical Transactions: Biological Scienc*es*, 364(1515), 301–311. [PubMed: 18977731]

Wright BA, Baese-Berk MM, Marrone N, & Bradlow AR (2015). Enhancing speech learning by combining task practice with periods of stimulus exposure without practice. The Journal of the Acoustical Society of America, 138(2), 928–937. [PubMed: 26328708]

Wright BA, Sabin AT, Zhang Y, Marrone N, & Fitzgerald MB (2010). Enhancing Perceptual Learning by Combining Practice with Periods of Additional Sensory Stimulation. Journal of Neuroscience, 30(38), 12868–12877. [PubMed: 20861390]

Yi HG, & Chandrasekaran B (2016). Auditory categories with separable decision boundaries are learned faster with full feedback than with minimal feedback. The Journal of the Acoustical Society of America, 140(2), 1332–1335. [PubMed: 27586759]

Yotsumoto Y, Chang L. h., Watanabe T, & Sasaki Y (2009). Interference and feature specificity in visual perceptual learning. Vision Research, 49(21), 2611–2623. [PubMed: 19665036]

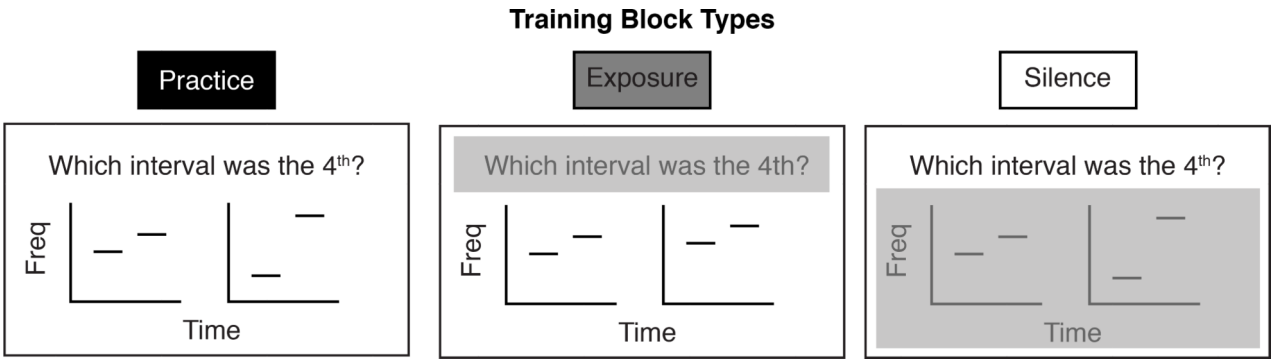

Figure 1.

Training block types. Schematics of a single trial for each of the three block types: During practice blocks (left panel), listeners were presented two musical-intervals and asked to indicate which of the two was a 4th. The starting pitch of each interval was selected at random, so the two intervals in each trial usually began on a different pitch. During exposure blocks (middle panel), listeners performed a non-auditory symbol-to-number matching task while in the background two examples of a 4th were presented on each "trial". During silence blocks (right panel), listeners were presented the same visual cues as during practice blocks, but without any sounds, and were asked to respond by making a "guess" for each trial. Listeners were trained using various combinations of the three block types.

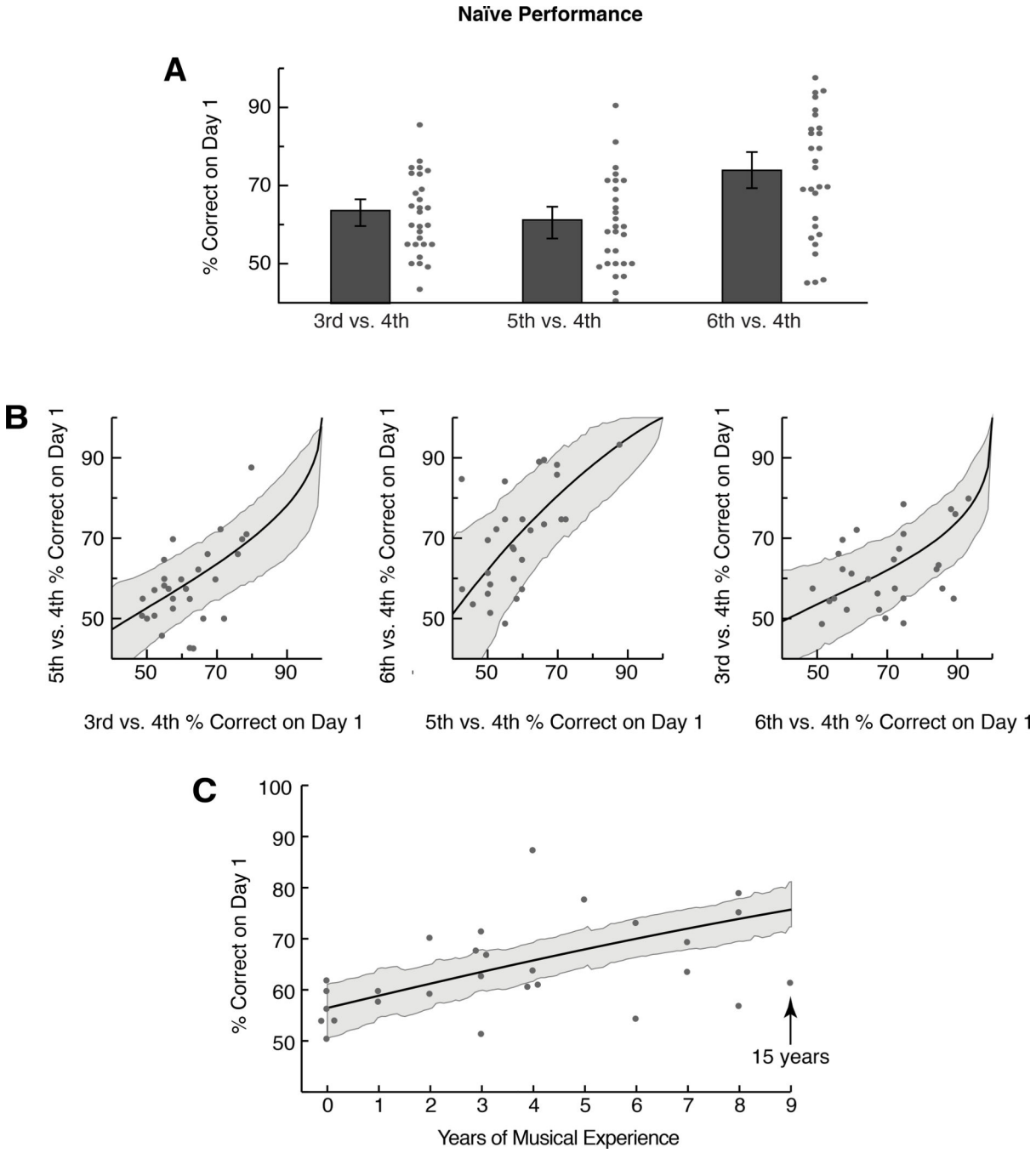

Figure 2.

Naïve performance (day 1). On each trial, listeners were presented a target musical interval (a 4th) and one of three foils (a 3rd, 5th or 6th) and asked to indicate which interval was the target. **A.** Individual response accuracy on day 1 separated by foil for 28 listeners (gray circles; raw data) with means and standard errors as estimated by a multi-level regression (bars and error bars). **B.** Individual response accuracy on day 1 (gray circles; raw data) for each of the three foils (x-axes) compared to another of the foils (y-axes) for the three pairwise comparisons (panels), with means and standard errors as estimated by the multi-

level regression (lines and gray regions). **C.** Individual response accuracy on day 1 across years of musical experience (gray circles; raw data) with the mean and standard error of this relationship as estimated by the multi-level regression (line and gray region).

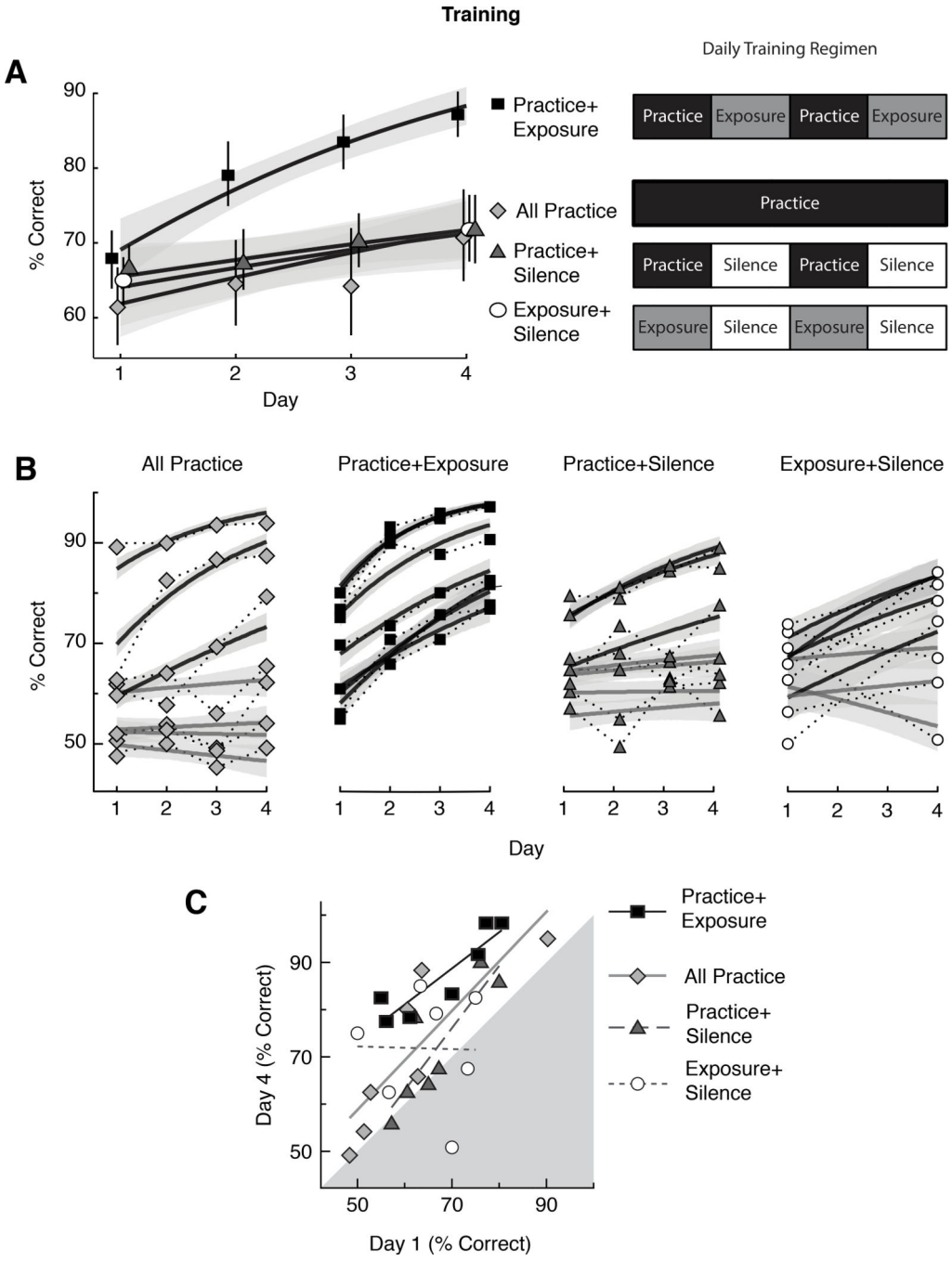

Figure 3.

Performance across days. **A.** Mean response accuracy (symbols and error bars; from raw data) across days (x-axis) on the trained musical-interval comparison task, and means and standard errors for those data as estimated by a single-level regression (lines and gray regions), for each of four training regimens (symbols). Each training session consisted of (1) continuous practice on the comparison task (All-Practice; n=7; gray diamond), (2) periods of practice alternating with periods of stimulus exposure in the absence of practice (Practice +Exposure; n=7; black square), (3) periods of practice alternating with periods of silence

(Practice+Silence; n=7; dark gray triangle), or (4) periods of stimulus exposure alone alternating with periods of silence (Practice+Exposure; n=7; white circle), all for the same total duration. **B.** Individual response accuracy (symbols and dotted lines; raw data) across days (x-axis) and training regimens (panels), with means and standard errors as estimated by a multilevel regression (solid lines and gray regions) for learners (black lines) and non-learners (gray lines). **C.** Individual mean response accuracy on day 1 (x-axis) and day 4 (y-axis) for each training regimen (symbols), and the fits by single-level regression for each group (lines).

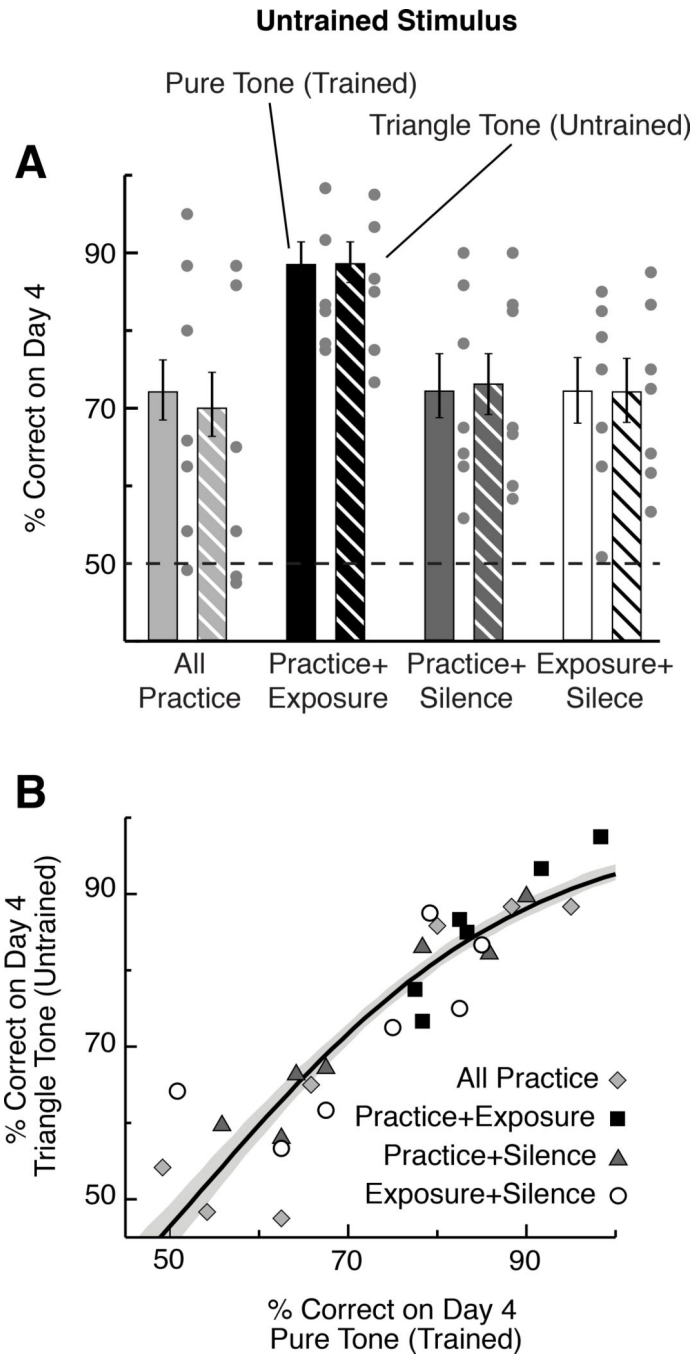

Figure 4.

Performance on the final day (day 4) for trained and untrained stimuli. **A.** Mean response accuracy on day 4 for the trained musical-interval comparison task with the trained pure-tone (solid bars) and untrained triangle-tone (striped bars) stimuli for each of four training regimens (shades) as estimated by a single-level regression, with standard errors from the regression (error bars) and individual data (gray circles; raw data). The dotted line indicates chance performance. **B.** Individual response accuracy on day 4 for pure-tone (x-axis) and triangle-tone (y-axis) stimuli, for each of the four training regimens (symbols; raw data),

with means and standard errors as estimated by a single-level regression (line and gray region).

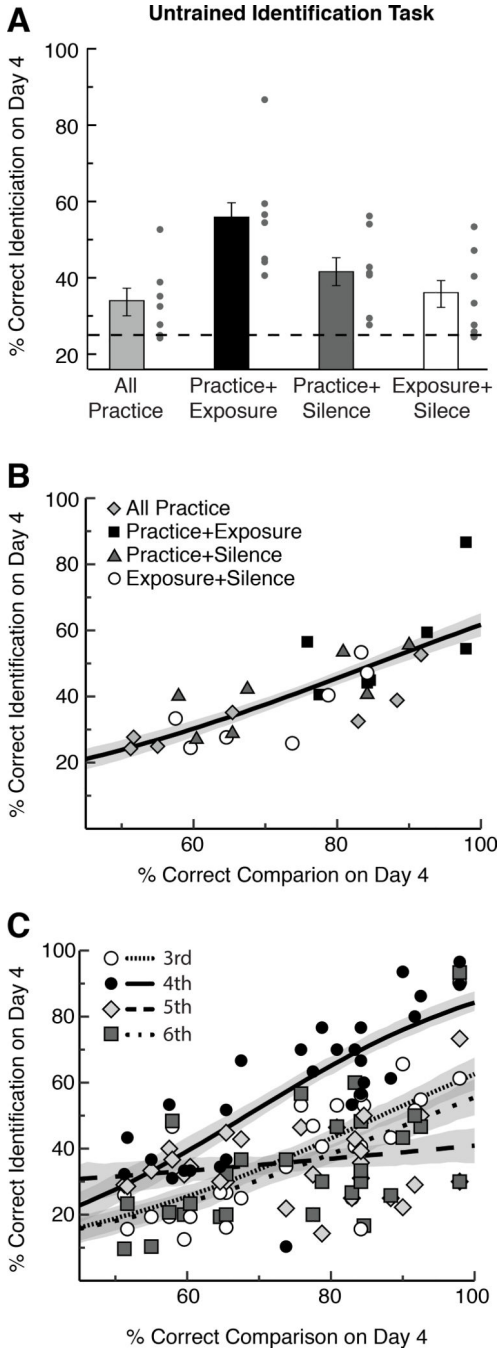

Figure 5.

Performance on the final day (day 4) for trained and untrained tasks. **A.** Mean response accuracy on day 4 for the untrained musical-interval comparison task (bars) for each of the four training regimens (shades) as estimated by a single-level regression, with standard errors from the regression (error bars) and individual data (gray circles; raw data). **B.** Individual response accuracy on day 4 for the trained musical-interval comparison task (x-axis) and the untrained comparison task (y-axis) for each of the four training regimens (symbols; raw data), with means and standard errors as estimated by a single-level

regression (line and gray region). **C.** Individual response accuracy on day 4 for the trained musical-interval comparison task (x-axis) and the untrained identification task (y-axis) for each interval (symbols), with means and standard errors as estimated by a single-level regression (lines and gray regions).