Published in final edited form as:
Curr Biol. 2017 December 04; 27(23): 3699–3705.e3. doi:10.1016/j.cub.2017.10.032.

Disruption of perceptual learning by a brief practice break

David F. Little[1,*], **Yu-Xuan Zhang**[2], and **Beverly A. Wright**[3]

[1]Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218

[2]State Key Laboratory of Cognitive Neuroscience and Learning, IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing 100875, China

[3]Department of Communication Sciences and Disorders, Knowles Hearing Center, Northwestern Institute for Neuroscience, Northwestern University, Evanston, IL 60208-3550

Summary

Some forms of associative learning require only a single experience to create a lasting memory [1,2]. In contrast, perceptual learning often requires extensive practice within a day for performance to improve across days [3,4]. This suggests that the requisite practice for durable perceptual learning is integrated throughout each day. If the total amount of daily practice is the only important variable, then a practice break within a day should not disrupt across-day improvement. To test this idea, we trained human listeners on an auditory frequency-discrimination task over multiple days and compared the performance of those who engaged in a single continuous practice session each day [4] with those who were given a 30-minute break halfway through each practice session. Continuous practice yielded significant perceptual learning [4]. In contrast, practice with a rest break led to no improvement, indicating that the integration process had decayed within 30 minutes. In a separate experiment, a 30-minute practice break also disrupted durable learning on a non-native phonetic classification task. These results suggest that practice trials are integrated up to a learning threshold within a transient memory store before they are sent en masse into a memory that lasts across days. Thus, the oft cited benefits of distributed over massed training [5,6] may arise from different mechanisms depending on whether the breaks occur before or after a learning threshold has been reached. Trial integration could serve as an early gate-keeper to plasticity, helping to ensure that longer lasting changes are only made when deemed worthwhile.

Results

Frequency discrimination

A 30-minute break midway through training disrupted across-day learning on a frequency-discrimination task (Figure 1). Four groups of young-adult, normal-hearing listeners were presented two pure-tone frequencies on each trial (1 kHz and < 1 kHz) and asked to select

[*]Correspondence: david.frank.little@gmail.com.
Lead Contact: David F. Little

Author Contributions: D.L., Y.Z. and B.W. designed the experiments. D.L. and Y.Z. collected the data. D.L. analyzed the data. B.W. and D.L. wrote the manuscript.

the lower frequency (Figure 1A). The measure of interest was the discrimination threshold—the frequency difference required for 79% correct performance. Listeners who received sufficient daily training with no break (long-daily-training group) (Figure 1B, 2nd column) improved across the training sessions and post-test (linear regression: $F(1,61) = 7.4$, $p = 0.008$) as well as between the pre- and post-tests (paired $t(7)= -6.3$, $p < 0.001$). They also improved more than controls (no training) as assessed at the post-test (ANCOVA: $F(1,15)=16.3$, $p = 0.001$). In contrast, listeners who received a 30-minute break halfway through each daily session of otherwise sufficient training (30-minute-break group) (Figure 1B, 3rd column) did not improve according to any of these measures (all $p \geq 0.203$). The lack of learning for the 30-minute-break group instead mirrored the result pattern for listeners who received only the (insufficient) amount of daily training that was provided before (or after) the break (short-daily-training group) (Figure 1B, 1st column) (all $p\geq 0.147$). Finally, listeners who received the same total number of training trials as the 30-minute break group, but with five 6-minute breaks equally spaced throughout each daily session (5×6-minute-break group) (Figure 1B, 4th column) improved across the training sessions and post-test ($F(1,61) = 5.1$, $p = 0.028$) as well as between the pre- and post-tests ($t(7)=3.7$, $p = 0.008$). The thresholds for this group did not differ from those of controls at the post-test ($F(1,15) = 1.6$, $p = 0.231$), but showed a trend in that direction on day 7 of training ($F(1,15) = 3.8$, $p = 0.070$).

Direct comparisons among the different trained groups revealed that the learning-curve slopes for the 30-minute-break and short-daily-training groups were shallower (less negative) than those for the long-daily-training and 5×6-minute-break groups [multiple linear regression, group by day: $F(4,242) = 16.1$, $p < 0.001$; contrasts by corrected multivariate T: long vs. short ($T(242) = 6.9$, $p < 0.001$), long vs. 30-minute ($T(242) = 4.4$, $p < 0.001$), 5×6-minute vs. short ($T(242) = 5.6$, $p < 0.001$), 5×6-minute vs. 30-minute ($T(242) = 2.8$, $p = 0.029$)]. The slopes for the 30-minute-break group were slightly steeper (more negative) than those for the short-daily-training group ($T(242)=2.811$, $p=0.028$), for whom thresholds actually increased slightly over days. The slopes did not differ significantly between the long-daily-training and 5×6-minute-break groups ($T(242)=1.56$, $p = 0.404$). In addition, when the learning-curve slopes were calculated across trials (rather than across days) and the analysis restricted to the first 5040 training trials—the total number presented to the 30-minute-break group—the slope for the 30-minute break group again was shallower than that for the long-daily-training and 5×6-minute-break groups [group by day: $F(4,1911) = 39.8$, $p < 0.001$; contrasts: 30-minute vs. long ($T(1911) = 7.4$, $p < 0.001$), 30-minute vs. 5×6-minute ($T(1911) = 5.5$, $p < 0.001$)].

As expected from the mean results, at the individual level, the post-test thresholds of the long-daily-training group (Figure 1C, 2nd column) and the day 7 thresholds of the 5×6-minute break group (Figure 1C, 4th column) were typically lower than the post-test thresholds of controls, regardless of the pre-test threshold, while the post-test thresholds of the 30-minute-break (Figure 1C, 3rd column) and short-daily-training (Figure 1C, 1st column) groups were interleaved with those of controls. Furthermore, there was significant improvement from day 1 of training to the post-test (by multi-level regression) for 6 of the 8 listeners in the long-daily-training group (Figure 1D, 2nd column) and 6 of 8 in the 5×6-minute-break group (to day 7; 0 of 8 to the post-test) (Figure 1D, 4th column), but only for 1 of 8 in the short-daily-training (Figure 1D, 1st column) and 30-minute break (Figure 1D, 3rd column) groups.

Within days (sessions), thresholds decreased significantly or nearly so for all three groups who received sufficient daily training [ANOVA of multiple linear regression, main effect of block: long-daily-training ($F(1,991) = 2.7$, $p < 0.001$), 5×6-minute break ($F(1,731) = 3.7$, $p < 0.001$), and 30-minute-break ($F(1,701) = 3.3$, $p = 0.052$)]. In addition, these three groups improved at similar rates within days in direct comparisons [main effect of block: $F(1,2428) = 10.7$, $p < 0.001$; no interaction of group by block: $F(2,2428) = 0.3$, $p = 0.527$; all pairwise comparison contrasts by multivariate T: $T(2428) \leq 1.5$, $p \geq 0.220$].

Non-native phonetic classification

A 30-minute break midway through training also disrupted learning of a non-native phonetic contrast (Figure 2). Monolingual, young-adult, normal-hearing speakers of American English were asked to classify the initial consonant of individual consonant-vowel syllables that varied along a voice-onset-time (VOT) continuum into one of three categories: positive VOT (labeled “pa”), near-zero VOT (labeled “ba”), and negative VOT (non-native to English; labeled “mba” for easy interpretation by listeners [8]) (Figure 2A). This three-way phonetic contrast occurs in many languages including Thai and Hindi, while there is only a two-way contrast between near-zero VOT and positive VOT in English. However, the three-way contrast can be acquired by native speakers of English with practice ([9,10]). The task was similar conceptually to a native speaker of Japanese learning to distinguish ‘r’ from ‘l.’ Feedback was provided after each trial throughout the single training session, but not during the post-test administered the next day (precluding simple comparisons between performance during the training on day 1 and the testing on day 2). The measure of interest was the slope for the non-native category boundary between negative (“mba”) and near-zero (“ba”) VOTs. At the post-test (Figure 2B), this slope differed from flat (no boundary; 0.5 on this scale) for listeners who practiced the classification task continuously during the training session (long-daily-training group) [$t(4) = 3.8$, $p = 0.018$; logistic regression: $p = 0.005$], but not for listeners who received a 30-minute break midway through the same amount of training (30-minute break group) [$t(4) = 1.8$, $p = 0.144$; logistic regression: $p = 0.268$]. The slope also differed between the two groups [Welch $t(4.9) = 3.0$, $p = 0.030$; logistic regression: $p = 0.022$].

Discussion

In order to persist across days, perceptual learning appears to require a sufficient amount of training per day [3,4], suggesting that multiple trials must integrate up to a learning threshold within a day to make a lasting memory. The lack of learning across days when there are too few trials within a day indicates that the sub-threshold content of the integrator can decay in less than 24 hours. The current results constrain that decay period to less than 30 minutes by showing—for two quite different tasks—that a 30-minute practice break midway through an otherwise sufficient amount of daily training can also prevent learning across days.

The idea that multiple trials must integrate within a day to form a durable memory comes from reports suggesting that perceptual learning persists across days only when there is sufficient training per day. This pattern has been documented previously for the same auditory frequency-discrimination task used here (> 360 trials/day required; [4]) and replicated for that task using three different variants of trial distribution [11]. It has also been described for a visual chevron-discrimination task (> 160 trials/day required [3]). The present data show this pattern yet again for the frequency-discrimination task, and extend the demonstration to a non-native phonetic classification task (>120 trials/day required), because the lack of across-day learning on these tasks when there was a 30-minute practice break indicates that neither the training period before nor the training period after the break was sufficient to yield lasting improvement. In an apparent exception to these results,1 to 5 training trials on day 1 led to small improvements in performance on day 2 on visual texture- and face-identification tasks [12], suggesting that the learning threshold for those tasks might be quite low or absent. If so, it may be that far fewer trials are necessary for learning when the stimuli to be compared differ along multiple dimensions rather than along a single dimension. However, while improvement for a given stimulus was observed on day 2 following only a few training trials with that stimulus on day 1, 21 different target stimuli were presented in visual noise on day 1 and on each trial the target stimulus was selected from 10 stimuli presented without visual noise. Thus, the aggregate of trials across different target stimuli or the aggregate of all of the stimuli presented on day 1 may have influenced the learning on day 2 in that investigation.

The indications that trial integration up to a learning threshold must occur within a day imply that the sub-threshold content of the integrator returns to baseline in less than 24 hours. The present results demonstrate that this decay can occur within just 30 minutes because a practice break of this length disrupted across-day learning on both a fine-grained discrimination task (pure-tone frequency discrimination) and a categorization task (non-native phonetic classification). They further suggest that the sub-threshold content of the integrator can persist for at least 6 minutes, because when the 30-minute break was split into five 6-minute breaks, across-day learning on the frequency-discrimination task was largely restored. We previously reported another quite different circumstance in which a practice break disrupted perceptual learning [11]. In that case, listeners practiced different tasks in the first and second halves of each session. When the two tasks were practiced in immediate succession, across-day learning on the first task (frequency discrimination) was enabled by training on the second (temporal-interval discrimination). That learning was reduced when the two practice periods were separated by a 15-minute break and abolished when the break was increased to 4 hours. While those data demonstrated a detrimental effect of a practice break, that outcome could have been due to the unusual two-task training regimen. The present results confirm that a practice break can disrupt across-day learning even when the same task is practiced before and after the break.

Where do the trials integrate?

Where do the trials integrate? The seemingly simplest possibility is that trial integration occurs within a transient memory store (Figure 3). According to this idea, if the learning threshold is reached within the time limit of the transient memory, the trials are sent *en* *masse* into a memory that lasts across days (Fig 3, middle column). Otherwise, the trials are erased (Figure 3, left and right columns). This transient memory would have to last at least multiple minutes, the time it takes for the threshold to be reached, suggesting that it differs from short-term auditory memory (echoic memory) [13–16] and working memory [17,18], which are typically estimated to last for several seconds. It could, however, reflect a trial-by-trial refreshing of one of these briefer memories. The idea that the learning processes engaged over the course of multiple trials during training are separate from those that ultimately enable a memory to last across days is consistent with other evidence from perceptual and motor learning. For example, events that disrupt across-day learning when introduced before the end of a training session can fail to do so when they occur immediately after the end of a training session, and vice versa, indicating that the processes operating during and after training are differentially vulnerable to the same intervening event [19–21]. In addition, learning within a training session, or the lack thereof, does not necessarily predict performance across sessions [4,22–24].

What about the trials is integrated?

What about the trials is integrated? Among many possibilities, the integrator could be simply counting trials (or stimulus presentations), with each contributing equally or by an amount modulated by the strength of permissive top-down signals. The integrator also could be computing a running average or cumulative prediction error across trials.

LTP and LTD

The idea that a distinct period of trial integration occurs prior to the formation of a durable perceptual memory has three intriguing counterparts with the induction of long-term potentiation (LTP) and long-term depression (LTD), synaptic changes thought to underlie at least some forms of long-term memory [25–27], including for perceptual learning [28–30]. First, just as trials must integrate up to a threshold to yield across-day learning, during the induction of LTP and LTD, more transient events must integrate up to a threshold to establish a more permanent state. Such thresholds exist for the transition from short-term plasticity to LTP and LTD [31] and for the transition from early to late LTP and LTD [32–34]. Second, just as the contents of the integrator can decay during a 30-minute break before reaching threshold, but can induce across-day learning once the threshold is reached, early LTP and LTD decay in less than 30 minutes [35] up to several hours [36], while late LTP and LTD can persist across days [37] (for review, see [32]). Third, just as learning processes active during a training session and those active following a training session can be disrupted by distinct events [19–21], early and late LTP and LTD can be differentially disrupted [36,38]. Thus, one possibility is that early LTP and LTD underlie trial integration and late LTP and LTD underlie the stabilization of trials across days. These three points repeat and add to a variety of other previously noted similarities between visual perceptual learning and LTP [39].

Implications for massed versus distributed training

Generally, distributed training—practice with breaks—is of greater benefit to learning than massed training—practice without breaks [5,6,40,41]. However, the present data demonstrate that a practice break can also disrupt learning. To the extent that this disruption reflects the interruption of an integration process leading to a learning threshold for across-day improvement, these data raise the possibility that the advantage typically observed with distributed training may be generated by two separate mechanisms, one that acts before and another that acts after this learning threshold has been reached. One mechanism may take advantage of practice breaks to facilitate integration up to the learning threshold. This mechanism could aid learning provided the breaks are not long enough for the contents of the integrator to decay (unlike in the present case). This possibility is consistent with evidence that molecular processes that underlie the benefit from multiple-minute breaks can precede those underlying long-term potentiation [42]. Another mechanism may rely on practice breaks to provide a needed rest during a post-threshold refractory period for trial integration, facilitating learning after the learning threshold has been reached. In this case, the breaks would be of benefit only if they were longer than the refractory period. This idea arises from reports that past some point additional training trials within the same daily training session do not enhance the amount of learning across days, suggesting that a learning process is saturated [4,40,41,43]. A further indication of the need for a reset stems from reports that extensive practice beyond that required to generate learning can even lead to across-day worsening [44,45]. It therefore appears that the learning benefits from practice breaks may arise from different mechanisms depending on the duration and timing of those breaks in relation to attainment of the learning threshold. A practical implication is that practice trials will be of no use if the practice breaks are too long prior to reaching the learning threshold (as reported here), or are too short after reaching that threshold. Thus, practice regimens could be optimized by selecting the lengths of practice breaks contingent on whether they occur before or after the learning threshold has been reached.

STAR Methods

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, David Little (david.frank.little@gmail.com).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Frequency Discrimination—Data are reported for 46 listeners (25 females) with a mean age of 21 years (standard deviation = 2.6). All had normal hearing and no previous experience with psychoacoustic tasks. All were paid for their participation. Excluded from this count was one listener whose log-transformed pre-test discrimination threshold for the trained condition was more than 2 standard deviations from the mean threshold of 133 listeners who participated in previous experiments (> 56 Hz, 1 listener; < 6.0 Hz, 0 listeners). All procedures were approved by the Northwestern University Office for the Protection of Research Subjects.

Non-native phonetic classification—Data are reported for 10 listeners (6 females) with a mean age of 18.7 years (standard deviation = 0.67). All were monolingual English speakers with normal hearing and were paid for their participation. All procedures were approved by the Northwestern University Office for the Protection of Research Subjects.

METHOD DETAILS

Frequency discrimination

Trained condition: The trained condition was pure-tone frequency discrimination with a standard frequency of 1 kHz (Figure 1A). Two digitally generated brief tones (15 ms, including 5-ms raised-cosine on/off ramps, 86 dB SPL) were presented in each presentation of a two-presentation, forced-choice trial. The two tones were separated by the same fixed temporal interval (t=100 ms onset to onset) in both presentations, but had a standard frequency (f=1 kHz) in one presentation and a lower comparison frequency (f - Δf) in the other. Listeners pressed a key on a computer keyboard to indicate which of the two randomly selected presentations contained the comparison sound (lower frequency). A visual display indicated whether the response was correct or incorrect after every trial throughout the experiment. The Δf value required for 79.4% correct performance, termed the discrimination threshold, was estimated using a 3-down/1-up adaptive procedure in 60-trial blocks [46]. Trial blocks that contained fewer than seven total reversals were excluded from analysis.

Training regimens: We compared frequency-discrimination thresholds from two new groups of listeners, trained on novel regimens (30-minute break and 5×6-minute break), to previously reported thresholds from two groups of trained listeners (short-daily-training and long-daily-training) [4,11] and a group of controls (no training) [11]. The previous data established two baselines: an amount of daily training that was insufficient to induce across-day learning (short daily training) and another amount that was sufficient (long daily training). All trained listeners participated in a single training session on each of 6-9 days. The number and temporal distribution of training trials in a session differed across the groups. In each session, the **30-minute-break** group (n=8) practiced 720 trials (~32 total minutes of practice) with a 30-minute break after the first 360 trials (new data). During the 30-minute breaks, listeners waited in a quiet environment outside of the testing booth under the supervision of the experimenter. They were allowed to participate in silent activities, but were not permitted to sleep. The **5×6-minute-break** group (n=8) practiced 720 trials with a 6-minute break after every 120 trials, for a total break duration of 30 minutes (5 breaks of 6 minutes) (new data). During the 6-minute breaks, listeners performed a written symbol-to-number matching task in silence inside of the testing booth. The **short-daily-training** group (n=8) practiced 360 trials per session (~16 minutes) with no break. The **long-daily-training** group (n=8) practiced 900 trials per session (~40 minutes) with no break [note that 720 stimulus presentations per session can yield as much learning across days as 900 practice trials [11]]. The **control** group (n=10) received no training during the training phase.

Pre- and post-tests: All listeners participated in a pre-test before and a post-test after the training phase. These tests comprised five threshold estimates (300 trials) on each of six conditions (1,800 total trials), including the trained condition. All trials of a given condition were presented contiguously. The condition order was randomized across listeners. Because our question focused on the direct influence of practice, we only report the results for the trained frequency-discrimination condition here. The remaining conditions employed either the frequency-discrimination task (described above) or a temporal-interval discrimination task [for details see [4,11]]. The pre- and post-tests were separated by an average of 14 days (standard deviation = 4.5) for the trained groups and 12 days (standard deviation = 2.5) for the control group. The pre-test followed a separate session during which listeners completed tone-detection tasks in quiet and in noise for ~1 h to familiarize them with the laboratory setting and the two-presentation, forced-choice procedure.

Non-native phonetic classification

Trained condition: The trained condition was phonetic classification of the initial consonant of individual consonant-vowel syllables. Each trial consisted of the presentation of a single syllable selected randomly from a 15-step voice-onset-time (VOT) continuum that ranged from −70 ms to +70 ms in 10-ms steps (Figure 2A). For each syllable, listeners used a computer mouse to select one of three category labels displayed on a computer screen: "mba" (< −25 ms VOT), "ba" (from −25 to 25 ms VOT), or "pa" (> 25 ms VOT). Trials were presented in blocks of 60. The stimuli were modified from tokens of "ba" and "pa" spoken by a female native speaker of American English. They were presented to the left ear at a comfortable listening level.

Training regimens: We compared performance between two groups of trained listeners (30-minute-break and long-training). Both groups participated in a single training session. The **30-minute-break** group (n= 5) practiced 240 trials (~25 minutes) with a 30-minute break after the first 120 trials. During the break, listeners waited either inside the testing booth or in a quiet environment outside of the testing booth, under the supervision of the experimenter. They were allowed to participate in silent activities, but were not permitted to sleep. The **long-training** group (n= 5) practiced 240 trials with no break. During training, the syllables were selected randomly from a trimodal distribution along the VOT continuum. The distribution peaks were centered on prototypical VOTs for each category: −50 ms for "mba," 0 ms for "ba," and +50 ms for "pa." Feedback was provided after each trial. Immediately before training, all listeners were given verbal instructions about the phonetic categorization task and samples of the prototypical tokens for the three categories.

Post-test: All listeners participated in a post-test the day after the training session. In the post-test, listeners completed 120 trials of the classification task in which each step from the VOT continuum was presented eight times. The syllables were presented in random order. No feedback on response accuracy was provided during the post-test.

The task, the stimuli, the stimulus distributions during training and the post-test, and the use of feedback were the same as in a previous investigation [47].

QUANTIFICATION AND STATISTICAL ANALYSIS

The outcomes for all statistical tests are reported in the results, including p-values. Outcomes with $p \leq 0.05$ were considered significant. Validation tests, to avoid violations of statistical modeling assumptions, are reported below as part of the description of each specific analysis. The degrees of freedom as reported in the results were determined by the number of listeners in each group (frequency discrimination: 8 for eachtrained group, 10 for controls; non-native: 5 for each group) and the number of measurements (days or blocks).

Figures show means and include error bars that indicate the standard error of the mean (SEM).

Frequency Discrimination—Unless otherwise specified, the dependent variable during analysis was the truncated mean of the log-transformed threshold estimates for each listener and day. After log-transformation, an individual threshold estimate from an individual listener was excluded from the daily mean estimate for that listener if that threshold was 3 standard deviations from the median of all individual threshold estimates for all listeners from the same day and training regimen. This approach excluded 16 threshold estimates, or 0.41% of the data.

Improvement in each trained group was assessed separately using a paired t-test of the pre-test and post-test thresholds, and a multiple linear regression of threshold (training days 1 through 7 and the post-test) on pre-test threshold and log-transformed day. The log scale for day yielded a better fit than a linear scale, consistent with the exponential form of the learning curve. In addition, post-test thresholds were compared between each trained group and controls using analysis of covariance (ANCOVA) with pre-test threshold as a covariate. Pre-test threshold was included as a covariate, because across all data analyzed there was a significant effect for this factor ($p < 0.001$).

Improvement across trained groups was compared using a multiple linear regression of the training-day (days 1-7) and post-test thresholds on pre-test threshold and the interaction of log-transformed day with training regimen. P-values for the difference between each pair of regression-line slopes were corrected for multiple comparisons [48]. To assess improvement on a per-trial rather than per-day basis, we followed the same multiple-linear-regression approach, but used the individual threshold estimates (excluding truncated values) and included the log of the trial number and its interaction with training regimen as predictors rather than the log-day predictors. Likewise, to assess within-session improvement we used the individual threshold estimates (without truncated values) and included both log-day and block and their interaction with training regimen as predictors. Suggesting that these models were appropriate for the data, variance appeared to be homogeneous across day (or trial as appropriate for the model) (Levene's Test, $p \geq 0.314$). Further, the 21-fold cross-validated root mean squared prediction error was ≤ 0.45 log Hz, which was between 0 and 6% smaller than the estimated prediction error of the full model.

Lastly, we evaluated improvement for each individual listener across all groups using a Bayesian multiple-level modeling analysis. For this analysis, we used each of the individual threshold estimates from an individual listener, excluding truncated values, instead of the daily mean threshold estimate for that individual. The model consisted of a set of coefficients for each individual and each training regimen. The coefficients in each set represented the individual days of training and the post-test (i.e., day was treated as a discrete variable). Pre-test threshold was included as a covariate in the prior mean of listener-level coefficients. Non-identifiable parameters were avoided by setting the regimen-level prior mean to zero. Parameters for the model hyperpriors, defined as in [49], were chosen to match the variation found in the pre-test thresholds of 133 listeners from prior experiments. This model was validated by posterior predictive checks which indicated that the model accurately predicted the root mean squared error and the 97.5%, 68.2%, 50%, 31.8% and 2.5% percentiles of the residuals with a standard error of ≤ 0.35 log Hz (posterior predictive p ≥ 0.112).

Model parameters were fitted to the data using the Stan modeling language (version 2.7.0) employing an MCMC algorithm [50] with 4 chains of 2,000 iterations, including 1,000 warmup iterations. Convergence of these four chains to the same solution was verified using the scale reduction statistic [51]. In all cases, the chosen number of samples was estimated to yield a prediction error of less than 0.01 log Hz (according to the standard error of the MCMC samples as determined by the effective sampling size).

Non-native phonetic classification—The measure of interest was the slope of the non-native category boundary between "mba" and "ba" at the post-test. To compute this slope, the classification data were fitted with a probit function separately for each listener. The slope of the function was scaled to a range from 0.0 to 1.0. A value of 0.0 indicated an abrupt, phonetically appropriate, shift from 100% to 0% "mba" identifications as the VOT became less negative, a value of 0.5 indicated a flat slope (no boundary), and a value of 1.0 indicated an abrupt, but phonetically inappropriate, shift from 0% to 100% "mba" identifications as the VOT became less negative. Only the post-test data were included in the analysis because of the differences between the training and testing periods in the VOT distributions (trimodal during training, uniform during testing) and feedback (feedback during training, no feedback during testing).

We compared performance on the post-test between the two trained groups using a Wald's t-test, which permits unequal variance, and a Bayesian robust logistic regression allowing for unequal variance. The logistic regression accounts for the fact that the dependent variable fell strictly within the range between 0 and 1, while the t-test assumes otherwise and may thus misestimate standard errors. In the regression, for numerical stability and robustness to outliers, the dependent variable (y) was transformed by r/2 + y(1-r)—where r was a relatively small proportion of the data range (0.01). Unequal variance across the groups was modelled with a separate beta distribution parameter. Priors over the coefficients of the logistic regression were zero-mean Cauchy distributions with a variance of 5 [52]. These choices were validated using posterior predictive checks which indicated that the resulting models accurately predicted the root mean squared error and the 97.5%, 68.2%, 50%, 31.8% and 2.5% percentiles of the residuals with a standard error of ≤ 0.02 units (posterior predictive p ≥ 0.524). Model parameters fitted using the same procedure as for the Bayesian analysis of the frequency-discrimination data yielded an estimated prediction error of ≤ 0.01 categorization slope units.

DATA AND SOFTWARE AVAILABILITY

The data collected and statistical analysis as reported in this manuscript are available at doi: 10.17632/nthkgdh8r2.1 along with instructions for reproducing the analyses on a new machine using the R and python programming environments.

Acknowledgments

This work was supported by the National Institute on Deafness and Other Communication Disorders–National Institutes of Health. We thank Dan H. Sanes and the anonymous reviewers for their helpful comments on the manuscript.

References

1. Tulving E. Episodic Memory: From Mind to Brain. Annu Rev Psychol. 2002; 53:1–25. [PubMed: 11752477]
2. Welzl H, D'Adamo P, Lipp HP. Conditioned taste aversion as a learning and memory paradigm. Behav Brain Res. 2001; 125:205–213. [PubMed: 11682112]
3. Aberg KC, Tartaglia EM, Herzog MH. Perceptual learning with Chevrons requires a minimal number of trials, transfers to untrained directions, but does not require sleep. Vision Res. 2009; 49:2087–2094. [PubMed: 19505495]
4. Wright BA, Sabin AT. Perceptual learning: How much daily training is enough? Exp Brain Res. 2007; 180:727–736. [PubMed: 17333009]
5. Cepeda NJ, Coburn N, Rohrer D, Wixted JT, Mozer MC, Pashler H. Optimizing Distributed Practice. Exp Psychol. 2009; 56:236–246. [PubMed: 19439395]
6. Donovan JJ, Radosevich DJ. A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. J Appl Psychol. 1999; 84:795–805.
7. Cohen, J. The concepts of power analysis. Hillsdale: Lawrence Erlbaum Assoc; 1988. Statistical power analysis for the behavioral sciences.
8. Tremblay K, Kraus N, Carrell TD, McGee T. Central auditory system plasticity: Generalization to novel stimuli following listening training. J Acoust Soc Am. 1997; 102:3762–3773. [PubMed: 9407668]
9. McClaskey CL, Pisoni DB, Carrell TD. Transfer of training of a new linguistic contrast in voicing. Percept Psychophys. 1983; 34:323–330. [PubMed: 6657433]
10. Pisoni DB, Aslin RN, Perey AJ, Hennessy BL. Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. J Exp Psychol-Hum Percept Perform. 1982; 8:297–314. [PubMed: 6461723]
11. Wright BA, Sabin AT, Zhang Y, Marrone N, Fitzgerald MB. Enhancing Perceptual Learning by Combining Practice with Periods of Additional Sensory Stimulation. J Neurosci. 2010; 30:12868–12877. [PubMed: 20861390]
12. Hussain Z, Sekuler AB, Bennett PJ. How much practice is needed to produce perceptual learning? Vision Res. 2009; 49:2624–2634. [PubMed: 19715714]
13. Demany, L., Semal, C. The Role of Memory in Auditory Perception. In: Yost, WA.Popper, AN., Fay, RR., editors. Auditory Perception of Sound Sources Springer handbook of auditory research. Springer; US: 2008. p. 77-113.
14. Mercer T, McKeown D. Decay uncovered in nonverbal short-term memory. Psychon Bull Rev. 2014; 21:128–135. [PubMed: 23801385]
15. Näätänen R, Paavilainen P, Rinne T, Alho K. The mismatch negativity (MMN) in basic research of central auditory processing: A review. Clin Neurophysiol. 2007; 118:2544–2590. [PubMed: 17931964]
16. Pasternak T, Greenlee MW. Working memory in primate sensory systems. Nat Rev Neurosci. 2005; 6:97–107. [PubMed: 15654324]
17. Baddeley A. Working Memory: Theories, Models, and Controversies. Annu Rev Psychol. 2012; 63:1–29. [PubMed: 21961947]
18. Ricker TJ, Cowan N. Differences between presentation methods in working memory procedures: A matter of working memory consolidation. J Exp Psychol Learn Mem Cogn. 2014; 40:417–428. [PubMed: 24059859]
19. Banai K, Ortiz JA, Oppenheimer JD, Wright BA. Learning two things at once: Differential constraints on the acquisition and consolidation of perceptual learning. Neuroscience. 2010; 165:436–444. [PubMed: 19883735]

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

21. Zach N, Kanarek N, Inbar D, Grinvald Y, Milestein T, Vaadia E. Segregation between acquisition and long-term memory in sensorimotor learning. European J Neurosci. 2005; 22:2357–2362. [PubMed: 16262674]

22. Huyck JJ, Wright BA. Late maturation of auditory perceptual learning. Dev Sci. 2011; 14:614–621. [PubMed: 21477199]

23. Huyck JJ, Wright BA. Learning, worsening, and generalization in response to auditory perceptual training during adolescence. J Acoust Soc Am. 2013; 134:1172–1182. [PubMed: 23927116]

24. Mednick SC, Nakayama K, Cantero JL, Atienza M, Levin AA, Pathak N, Stickgold R. The restorative effect of naps on perceptual deterioration. Nat Neurosci. 2002; 5:677–681. [PubMed: 12032542]

25. Bliss TVP, Collingridge GL. A synaptic model of memory: Long-term potentiation in the hippocampus. Nature. 1993; 361:31–39. [PubMed: 8421494]

26. Kemp A, Manahan-Vaughan D. Hippocampal long-term depression: Master or minion in declarative memory processes? Trends Neurosci. 2007; 30:111–118. [PubMed: 17234277]

27. Rioult-Pedotti MS, Friedman D, Donoghue JP. Learning-Induced LTP in Neocortex. Science. 2000; 290:533–536. [PubMed: 11039938]

28. Sale A, De Pasquale R, Bonaccorsi J, Pietra G, Olivieri D, Berardi N, Maffei L. Visual perceptual learning induces long-term potentiation in the visual cortex. Neuroscience. 2011; 172:219–225. [PubMed: 21056088]

29. Beste C, Wascher E, Güntürkün O, Dinse HR. Improvement and Impairment of Visually Guided Behavior through LTP- and LTD-like Exposure-Based Visual Learning. Curr Biol. 2011; 21:876–882. [PubMed: 21549600]

30. Ragert P, Kalisch T, Bliem B, Franzkowiak S, Dinse HR. Differential effects of tactile high- and low-frequency stimulation on tactile discrimination in human subjects. BMC Neurosci. 2008; 9:9. [PubMed: 18215277]

31. Lisman J, Yasuda R, Raghavachari S. Mechanisms of CaMKII action in long-term potentiation. Nat Rev Neurosci. 2012; 13:169–182. [PubMed: 22334212]

32. Abraham WC. How long will long-term potentiation last? Philos Trans R Soc Lond B Biol Sci. 2003; 358:735–744. [PubMed: 12740120]

33. Hulme SR, Jones OD, Abraham WC. Emerging roles of metaplasticity in behaviour and disease. Trends Neurosci. 2013; 36:353–362. [PubMed: 23602195]

34. Redondo RL, Morris RGM. Making memories last: The synaptic tagging and capture hypothesis. Nat Rev Neurosci. 2011; 12:17–30. [PubMed: 21170072]

35. Seidenbecher T, Reymann KG, Balschun D. A post-tetanic time window for the reinforcement of long-term potentiation by appetitive and aversive stimuli. Proc Natl Acad Sci USA. 1997; 94:1494–1499. [PubMed: 9037081]

36. Krug M, Lössner B, Ott T. Anisomycin blocks the late phase of long-term potentiation in the dentate gyrus of freely moving rats. Brain Res Bull. 1984; 13:39–42. [PubMed: 6089972]

37. Abraham WC, Mason SE, Demmer J, Williams JM, Richardson CL, Tate WP, Lawlor PA, Dragunow M. Correlations between immediate early gene induction and the persistence of long-term potentiation. Neuroscience. 1993; 56:717–727. [PubMed: 8255430]

38. Sajikumar S, Frey JU. Resetting of “synaptic tags” is time- and activity-dependent in rat hippocampal CA1in vitro. Neuroscience. 2004; 129:503–507. [PubMed: 15501607]

39. Aberg KC, Herzog MH. About similar characteristics of visual perceptual learning and LTP. Vision Res. 2012; 61:100–106. [PubMed: 22289647]

40. Molloy K, Moore DR, Sohoglu E, Amitay S. Less is more: Latent learning is maximized by shorter training sessions in auditory perceptual learning. PLoS ONE. 2012; 7:e36929. [PubMed: 22606309]

41. Ofen-Noy N, Dudai Y, Karni A. Skill learning in mirror reading: How repetition determines acquisition. Brain Res Cogn Brain Res. 2003; 17:507–521. [PubMed: 12880920]

42. Naqib F, Farah CA, Pack CC, Sossin WS. The Rates of Protein Synthesis and Degradation Account for the Differential Response of Neurons to Spaced and Massed Training Protocols. PLOS Comput Biol. 2011; 7:e1002324. [PubMed: 22219722]

43. Ortiz JA, Wright BA. Differential rates of consolidation of conceptual and stimulus learning following training on an auditory skill. Exp Brain Res. 2010; 201:441–451. [PubMed: 19902196]

44. Censor N, Sagi D. Benefits of efficient consolidation: Short training enables long-term resistance to perceptual adaptation induced by intensive testing. Vision Res. 2008; 48:970–977. [PubMed: 18295817]

45. Censor N, Sagi D. Global resistance to local perceptual adaptation in texture discrimination. Vision Res. 2009; 49:2550–2556. [PubMed: 19336239]

46. Levitt H. Transformed Up-Down Methods in Psychoacoustics. J Acoust Soc Am. 1971; 49:467–477.

47. Wright BA, Baese-Berk MM, Marrone N, Bradlow AR. Enhancing speech learning by combining task practice with periods of stimulus exposure without practice. J Acoust Soc Am. 2015; 138:928–937. [PubMed: 26328708]

48. Hothorn T, Bretz F, Westfall P. Simultaneous inference in general parametric models. Biom J. 2008; 50:346–363. [PubMed: 18481363]

49. Gelman A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Anal. 2006; 1:515–534.

50. Hoffman MD, Gelman A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. J Mach Learn Res. 2014; 15:1593–1623.

51. Gelman, A., Carlin, JB., Stern, HS., Rubin, DB. Bayesian Data Analysis. Taylor & Francis; 2014.

52. Gelman A, Jakulin A, Pittau MG, Su YS. A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models. Ann Appl Stat. 2008; 2:1360–1383.

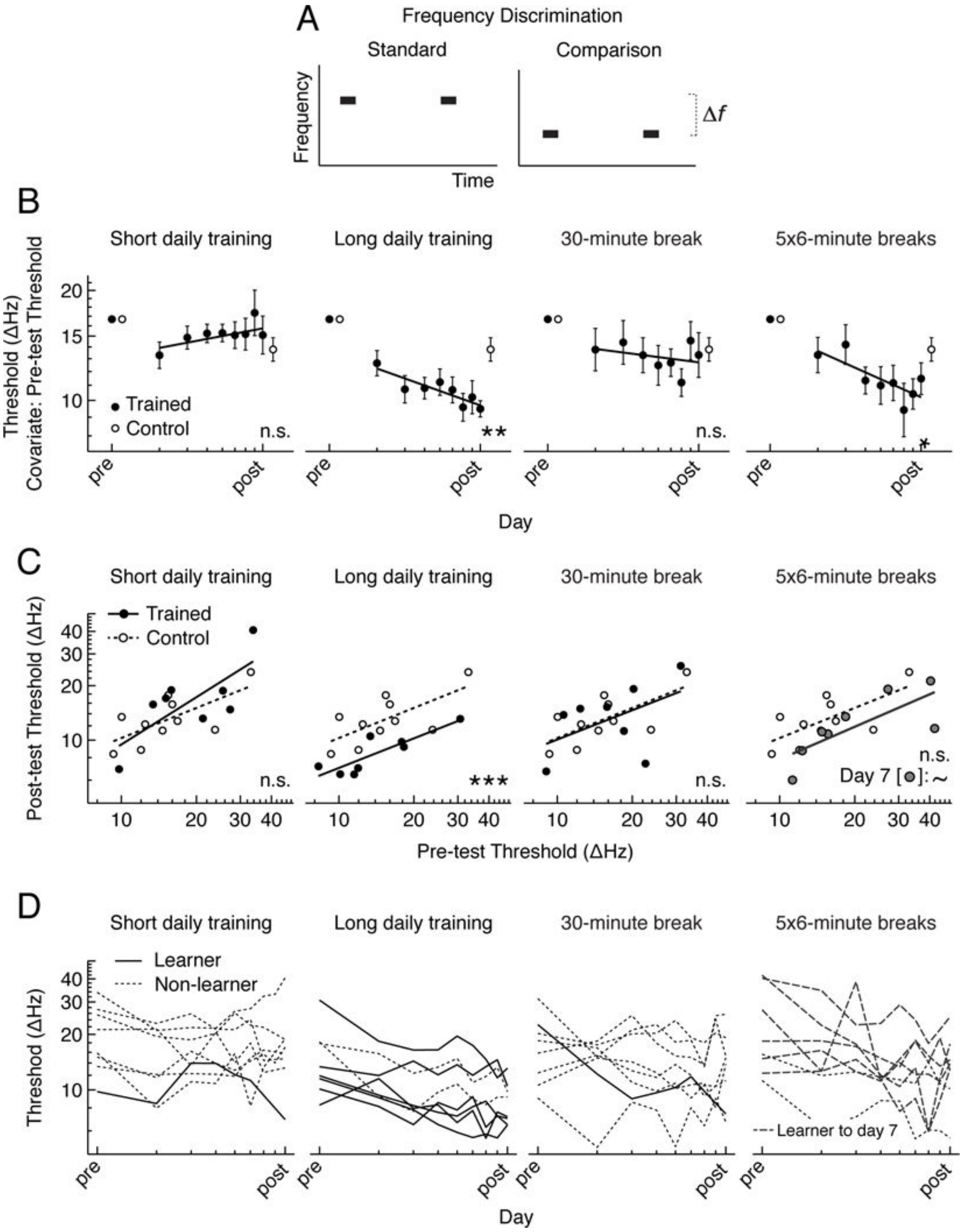

Figure 1. Frequency discrimination

A. Frequency-discrimination task. **B** Group mean frequency-discrimination thresholds for the trained groups (n=8 per group; blackcircles) and controls (n=10; open circles) for each of the four training regimens (columns). Thresholds across days are adjusted using pre-test threshold as a covariate [7] and fitted with least-squares regression lines across the log of day number. Axis scales are in log units of day (x-axis) and frequency (y-axis). Error bars indicate +/− SEM. Asterisks denote significant improvement across the training sessions and post-test, as well as between the pre- and post-tests, for trained listeners (p ≤ 0.008). **C.**

Individual pre-test (x-axis) versus post-test (y-axis) thresholds (symbols) on a log scale fitted with least-squares regression lines separately for trained listeners (solid lines; black circles; gray circles for day 7 thresholds) and controls (dashed lines; open circles). Asterisks denote significantly lower post-test thresholds for trained listeners than controls, using pre-test threshold as a covariate ($p = 0.001$). The tilde ("~") indicates a trend for significance for day 7 thresholds vs. pre-test thresholds ($p = 0.070$). **D.** Individual frequency-discrimination thresholds across days for each of the four trained groups. Axis scales are in log units. Solid lines indicate significant learning from day 1 to the post-test ($p \leq 0.05$), dashed lines significant learning from day 1 to day 7, and dotted lines no significant learning indicate a non-significant result.

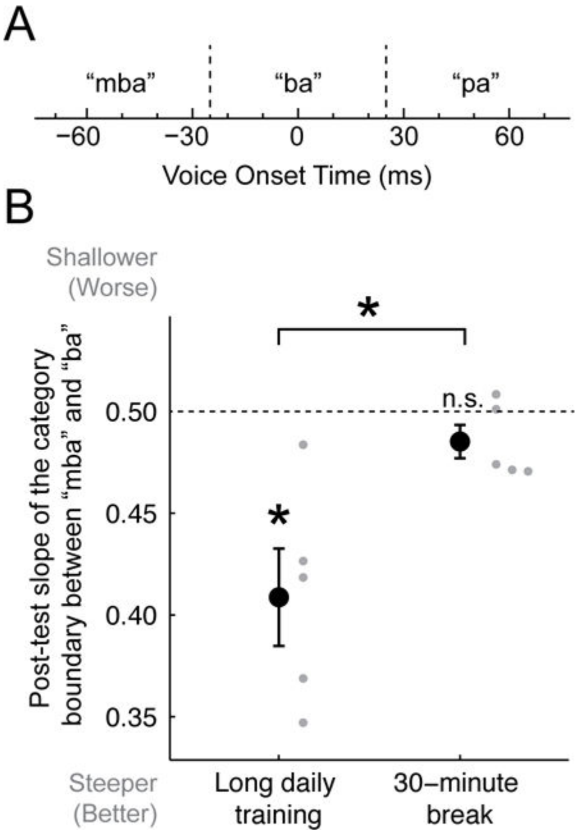

Figure 2. Non-native phonetic classification

A. Task stimuli. Each tick mark designates the voice-onset-time of one of the 15 different consonant-vowel stimuli. Dotted lines demarcate the category boundaries for the non-native ("mba" vs. "ba") and native ("ba" vs. "pa") phonetic contrasts, as indicated by the feedback provided during training on day 1. **B.** Mean (black circles) and individual (gray circles) slopes for the non-native category boundary between negative ("mba") and near zero ("ba") VOTs at the post-test assessed without feedback on day 2 for the two trained groups (n=5 per group). The slope of the function is scaled such that the closer the value is to zero, the sharper the category boundary. Error bars indicate +/− SEM. Asterisks denote a significant difference between the post-test and chance performance, and between the post-test of the two trained groups ($p \leq 0.030$).

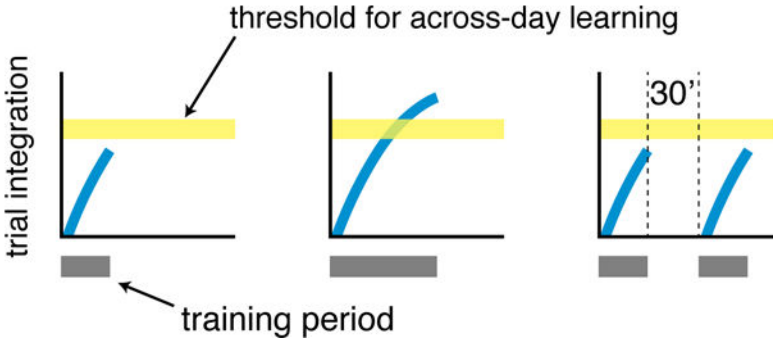

Figure 3. Schematic diagram of the proposed trial integration process

Practice trials (dark gray rectangles) integrate (thick black line) in a transient memory store and only stabilize in a store that lasts across days when integration surpasses a learning threshold (light gray bar). Given insufficient training within the transient-memory period the trials do not persist across days (left panel)), but with sufficient training the trials are retained (middle panel). According to this idea, the 30-minute break disrupts learning because the transient memory has largely reset during this break (right panel).