Attention, Perception, & Psychophysics (2023) 85:525–542
https://doi.org/10.3758/s13414-023-02653-0

# Generalizing across tonal context, timbre, and octave in rapid absolute pitch training

Noah R. Bongiovanni[1] · Shannon L.M. Heald[2] · Howard C. Nusbaum[2] · Stephen C. Van Hedger[3,4]

Accepted: 3 January 2023 / Published online: 23 January 2023


## Abstract

Absolute pitch (AP) is the rare ability to name any musical note without the use of a reference note. Given that genuine AP representations are based on the identification of isolated notes by their tone *chroma*, they are considered to be invariant to (1) surrounding tonal context, (2) changes in instrumental timbre, and (3) changes in octave register. However, there is considerable variability in the literature in terms of how AP is trained and tested along these dimensions, making recent claims about AP learning difficult to assess. Here, we examined the effect of tonal context on participant success with a single-note identification training paradigm, including how learning generalized to an untested instrument and octave. We found that participants were able to rapidly learn to distinguish C from other notes, with and without feedback and regardless of the tonal context in which C was presented. Participants were also able to partly generalize this skill to an untrained instrument. However, participants displayed the weakest generalization in recognizing C in a higher octave. The results indicate that participants were likely attending to pitch height in addition to pitch chroma – a conjecture that was supported by analyzing the pattern of response errors. These findings highlight the complex nature of note representation in AP, which requires note identification across contexts, going beyond the simple storage of a note fundamental. The importance of standardizing testing that spans both timbre and octave in assessing AP and further implications on past literature and future work are discussed.

**Keywords** Absolute pitch · Training · Generalization · Octave · Timbre

## Introduction

Absolute pitch (AP), sometimes referred to as "perfect pitch," is the ability to name or produce a musical note without the use of a reference (e.g., for reviews, see Deutsch, 2013; Levitin & Rogers, 2005; Takeuchi & Hulse, 1993). AP is a rare ability, estimated to be present in approximately one in 10,000 people (Bachem, 1955). The exact base rate in the population is currently debated and may depend on factors such as cultural upbringing and early linguistic experience (Deutsch et al., 2009; Miyazaki et al., 2012). Understanding the base rate of AP also depends largely on awareness of how the ability is operationalized; for example, Van Hedger et al. (2020) found more continuous performance when using a measure that simultaneously incorporated response speed and absolute deviation compared to a measure that simply coded each response as correct or incorrect.

The mechanisms underlying AP acquisition are still uncertain and remain a subject of interest. The most widely accepted theory suggests that the skill can only be learned during a developmental window in early childhood (*critical period theory*) (Levitin & Rogers, 2005; Levitin & Zatorre, 2003). The critical period theory of AP is supported in principle by several converging findings. First, there is an association between beginning musical instruction at an early age and AP incidence (Baharloo et al., 2000; Deutsch et al., 2006; Gregersen et al., 1999), although early musical training may not be necessary or sufficient to develop AP (Brown et al., 2002). Second, pharmacological interventions in adults, hypothesized to "reopen" a critical period of learning, have been associated with improved AP learning (Gervain et al., 2013). Third, explicit learning studies have shown that children

✉ Noah R. Bongiovanni
nbongiov@nd.edu

[1] Department of Music, University of Notre Dame, Notre Dame, IN, USA

[2] Department of Psychology, University of Chicago, Chicago, IL, USA

[3] Department of Psychology, Huron University College at Western, London, ON, Canada

[4] Western Institute for Neuroscience, Western University, London, ON, Canada

perform significantly better than adults and adolescents when trained to identify a single target note (Crozier, 1997; Russo et al., 2003). Fourth, adult AP training studies generally find only modest improvements in note identification (Cuddy, 1968; Hartman, 1954; Lundin, 1963), though more recent adult training studies have suggested that explicit training in some adults may lead to genuine-like AP under some circumstances (Van Hedger et al., 2019; Wong, Lui, et al., 2020a; Wong, Ngan, et al., 2020b). Overall, the existing research suggests that success in learning AP in adulthood seems dependent on individual differences, and acquiring AP to a degree indistinguishable from a "genuine" AP possessor, if possible, is quite difficult and requires consistent, explicit training.

## The role of context in absolute pitch (AP) representations

When considering the trainability of AP across the lifespan, it is important to first describe the nature of AP representations as observed among genuine AP possessors. Without careful consideration of genuine AP representations, it is possible to observe learning that, on the surface, might appear to be genuine AP but upon further examination can be explained via alternative mechanisms.

AP representations are based on pitch class, or *chroma* (cf. pitch height; Warren et al., 2003). Pitch chroma is the quality that makes all notes separated by octave relationships (a 2:1 frequency ratio) sound similar (e.g., Takeuchi & Hulse, 1993). For example, to a genuine AP possessor, the high C note of a piccolo and the low C note of a tuba would retain the same qualia in terms of chroma, despite being clearly differentiable in terms of pitch height and instrumental timbre. Moreover, to an AP possessor, the Cs played on both a piccolo and tuba would be just as identifiable if they were played in isolation compared to if they were heard in a melodic context. The remaining paragraphs of this section further unravel these critical features of AP, which are all incorporated into the present experimental paradigm.

First, AP is characterized by representations of musical notes that are putatively invariant to the tonal context in which a note is heard (Takeuchi & Hulse, 1993). As alluded to in the previous paragraph, among AP possessors, a "C" retains its qualia (i.e., its "C-ness") regardless of whether it is heard in isolation, heard in a context that would make it sound congruent with surrounding notes (e.g., C major), or heard in a context that would make it sound incongruent with surrounding notes (e.g., F# major). This observation has led researchers to compare AP representations to visual color representations (Levitin & Rogers, 2005), which also retain properties of constancy under a variety of lighting contexts (e.g., see Kraft & Brainard, 1999) – for example, the "redness" of an apple under both broadband sunlight and fluorescent grocery store light. Although both color and AP representations can be influenced by immediate context (e.g., Lotto & Purves, 2000; Van Hedger et al., 2018b), which is likely reflective of broader contextual influences on perception (e.g., Chambers et al., 2017; Nguyen & McKendrick, 2016), it is generally the case that an AP possessor will have little trouble identifying a given note (e.g., C) regardless of tonal context.

In contrast, tonal context exerts large influences on both the perception and memory of pitch among non-AP possessors, reflective of a greater reliance on *relative* pitch processing (e.g., Takeuchi & Hulse, 1993). Outside of experimental contexts, musical pitches are generally not heard in isolation. Instead, they are heard in relation to other notes, specifically, the set of notes within which a piece of music is organized, such as a key. Any note can be described in context as either diatonic (i.e., part of the same subset of notes or key signature) or non-diatonic (i.e., not part of the same subset of notes or key signature). Critically, how the note is perceived is therefore contextually determined – the same pitch chroma (e.g., C) may be heard as highly congruent (e.g., in C major) or highly incongruent (e.g., in F# major). This poses a potential challenge for forming a long-term chroma representation based on pitch, particularly when the surrounding tonal context appears to "color" chroma perception.

The notion that key contexts affect chroma perception is indeed evidenced by a rich literature demonstrating the connection between tonal context and the perception and memory of pitch (e.g., Deutsch, 1972a, 1972b, 1982; Deutsch & Roll, 1974; Dewar et al., 1977; Krumhansl, 1979; Krumhansl & Castellano, 1983). For example, Krumhansl (1979) concluded that tonal context has a strong effect on the goodness-of-fit ratings listeners assign to probe-tones, and demonstrated that diatonic tones were more readily remembered by listeners than non-diatonic tones. Another more recent study demonstrated that tonal expectations can influence aspects of lower-level perceptual judgments of pitch, including judgments of (mis)tuning (Marmel et al., 2008). Although a strong sense of tonal context can be eliminated in experimental contexts by presenting a chromatic context (i.e., presenting all 12 notes with equal probability), which is often the approach used in AP assessments, doing so is not an ecologically valid representation of how tonal music is experienced outside of research contexts. This may also explain why some studies show limited generalizability in identifying notes outside of non-musical training contexts. Thus, the present study also aimed to assess how different tonal contexts influence the explicit training of AP, which will aid in the understanding of AP as a perceptual skill, as well as the importance of context for note perception generally.

Second, pitch chroma representations are dissociable from both instrumental timbre and octave. As such, a hallmark of genuine AP is the ability to name *any* stable pitched sound in terms of its musical note name, regardless of instrument or

octave register (e.g., for a review, see Takeuchi & Hulse, 1993). Consequently, AP training studies must assess whether learning can generalize across octave and timbre to make claims about the trainability of AP categories. Indeed, an inability to generalize across timbre or octave is commonly treated as distinct from AP, being labeled as "pseudo" or "quasi" AP (Bachem, 1937). An extreme example of limited generalizability across timbres is the phenomenon of *instrument-specific AP*, which describes subjects who display AP-like performance for their primary instrument but lack the ability to generalize their skills to instruments of a different timbre (Reymore & Hansen, 2020). It is suggested that for such individuals, timbral cues or even motor imagery play an essential part in their ability to identify pitches; as such, they do not have genuine AP, which primarily relies on chroma cues.

However, this is not to say that AP possessors do not make use of cues beyond pitch chroma recognition; indeed, ample evidence suggests the contrary. It is demonstrated that AP possessors have less success with identifying sine tones compared to complex instrumental timbres (Lockhead & Byrd, 1981; Miyazaki, 1989), as well as difficulty identifying complex instrumental timbres relative to common musical timbres (Miyazaki, 1989). AP possessors also have trouble identifying pitches produced by natural or synthesized vocal tones compared to non-vocal tones (Vanzella & Schellenberg, 2010). Another study determined that violinists with "genuine" AP still demonstrated greater accuracy at tuning a violin pitch to 440 Hz than when asked to do so with a clarinet pitch (Brammer, 1951). This aligns with Sergeant (1969), who concluded that musicians with and without AP both name pitch most accurately on their instruments of greatest and/or earliest exposure. AP possessors also respond significantly slower when identifying a target note from a series when the series contains multiple timbres, even though timbre is irrelevant to the categorization task (Van Hedger, Heald, & Nusbaum, 2015b). These experiments suggest that AP possessors must be using other auditory dimensions in addition to pitch chroma for pitch identification, but the degree of dependence on these features varies situationally.

Given that the literature has argued that genuine AP is based on the categorization of pitch chroma and not pitch height (Bachem, 1955; Kim & Knösche, 2016, 2017), it is critical that AP training studies demonstrate that any learning generalizes beyond the trained octave. Without demonstrating an ability to label notes beyond the trained octave, it is unclear if the learner has relied on the use of pitch *height*, instead of forming explicit categories based on pitch *chroma*. Indeed, given that AP possessors have been shown to make frequent octave errors (e.g., mistaking an A4 (440 Hz) for an A3 (220 Hz); cf. Kim & Knösche, 2016; Miyazaki, 1988; Takeuchi & Hulse, 1993), there is little debate that AP is defined on the basis of pitch *chroma* and not *height*. Yet, these dissociations of pitch *chroma* and *height* do not mean that AP possessors' pitch labeling ability is not sensitive to octave change. Notably, genuine AP possessors see decreased accuracy for notes from extremely low or high octaves (Oxenham, 2012). Furthermore, AP possessors are slower to respond when making note category judgments in situations when octaves can vary – even when the two octaves presented are in a "comfortable" middle range (Van Hedger, Heald, & Nusbaum, 2015b). Together, these studies indicate that while genuine AP is marked by robust note-labeling performance across multiple octaves, there is some evidence of discontinuity of this performance due to octave change and extrema.

Despite agreement that the skill of AP is (1) invariant to different tonal contexts, (2) generalizes across instrumental timbre, and (3) generalizes across octaves, prior research examining AP has unfortunately used inconsistent approaches in measuring generalization along these dimensions. These inconsistencies may have contributed to the mixed and sometimes conflicting results with respect to explicit AP training across the lifespan. For example, some of the foundational empirical studies used to support critical periods for AP have focused on training a single frequency (e.g., C4 or A4) and did not assess either timbre or octave generalization post-training (Crozier, 1997; Russo et al., 2003). In Gervain et al.'s (2013) proof-of-concept demonstration that valproate can reopen a critical period for learning AP, multiple octaves are tested but only a single timbre (piano) is used. In recent investigations of adult AP training, some post-training assessments have incorporated multiple timbres and octaves (Van Hedger et al., 2019; Van Hedger et al., 2015a; Wong, Lui, et al., 2020a), whereas others have only tested a single timbre and octave (Wong, Ngan, et al., 2020b). Establishing consistent testing standards, including robust assessments that span across timbres and octaves, is important in creating a more coherent picture of AP performance. By ensuring that the appropriate psychological construct is being measured, researchers can more appropriately assess the multidimensionality of AP, including investigating how AP might utilize more general cognitive processes used in other perceptual abilities. For this reason, we additionally assessed learning in our AP training study using multiple post-tests that specifically measured how learning generalized to a novel timbre and to a novel octave.

## The present experiment

The present study examines the influence of tonal context on the explicit training and generalization of AP. Adult participants were asked to recognize a target note – the "middle C" (C4) of a piano – from a series of non-target notes. The relative proportion of target notes was high (50% of all heard notes) at the beginning of training and gradually decreased throughout training, conceptually similar to other AP training paradigms

(Brady, 1970; Cuddy, 1968). Moreover, the approach of training a single note category as proof-of-concept AP has been adopted by prior research, including studies that have been used to empirically support critical periods in AP acquisition (Crozier, 1997; Russo et al., 2003).

Unlike prior research, however, the present experiment systematically manipulated the tonal context of training (i.e., the tonal relationship between the trained and untrained notes) to determine whether the tonal context (diatonic, non-diatonic, or mixed) influenced the efficacy of learning C. Additionally, the present study varied three auditory dimensions in testing to assess different aspects of how learning generalized beyond the specific conditions of training. First, *chromatic generalization* involved testing participants in a *chromatic* context (in which all 12 notes were equiprobable), assessing the extent to which tonal training could be applied to an atonal context. Second, *timbre generalization* used notes from a timbre (French horn) that was not experienced in training. Third, *octave generalization* used notes from an octave range that was not experienced in training (making the target note C5).

Based on the findings of Brady (1970), who reported considerable success in AP learning when adopting a "fixed scale" strategy (e.g., attempting to hear all notes in the context of a single key), we hypothesized that contexts in which C4 always sounded congruent (diatonic condition) or incongruent (non-diatonic condition) might elicit stronger AP learning than a context in which C4 could either sound congruent or incongruent on a trial-by-trial basis (mixed condition). However, we additionally hypothesized that the mixed condition training might lead to stronger generalization, as it would provide participants with the most varied context in which the target note could be experienced (e.g., see Ahissar & Hochstein, 2004). In contrast, participants in the diatonic and non-diatonic conditions could potentially use heuristics (e.g., that the target note sounded "good" or "bad") that would serve them well when being tested in an identical context as training but would not be effective once the tonal context was changed relative to training. Additionally, based on prior work (e.g., Van Hedger et al., 2015a, b), we hypothesized that performance in the timbre and octave generalization tests would be attenuated relative to the specific test, which mirrored the exact conditions of training.

## Method

### Participants

A total of 177 participants successfully completed the experiment. Participants were recruited through Amazon Mechanical Turk, using the Cloud Research platform (Litman et al., 2017). Cloud Research provides additional participant recruitment options meant to facilitate high-quality online data collection. Specifically, we only recruited participants who had previously passed attention checks administered by Cloud Research. Participants were not specifically recruited based on their musical background or auditory working memory ability (cf. Van Hedger et al., 2019). Due to a programming error, several participants' data from the AP training and testing portion of the experiment were not able to be successfully linked with the demographic and musical background questionnaire. Thus, reports of demographic or musical training variables use a subset ($n = 90$) of participants for whom this information could be recovered.

The sample size was set primarily based on the availability of funds. However, it should be noted that the current sample size is adequately powered (1 − = .84) to detect medium effect sizes ($f = 0.25$) with a between-participant design containing three groups. Participants were well beyond a developmental stage in which a critical period would be presumed to be open ($M = 41.18$ years old, $SD = 11.61$ years, range: 20–72 years old). A minority (33.3%) of participants reported that they had either played a musical instrument or sang. All participants provided informed consent in accordance with the Declaration of Helsinki and were compensated US$7.50 for completing the experiment.

### Materials

The AP training and testing script was programmed in jsPsych (de Leeuw, 2015). The follow-up questionnaire was administered in Qualtrics (Provo, UT). Participants accessed the experiment from their own devices.

All audio samples used in the AP training and testing procedure were played on a Yamaha Portable Grand Piano (DGX-640) keyboard using preloaded patches. The piano samples came from the "001 Live! Grand Piano" patch, whereas the French horn samples came from the "103 French Horn" patch. Touch and dynamics were standardized so that no note would be louder upon attack than another (rate of decay differs naturally based on the pitch of the note/length of the "string" for the piano samples, though the effect is minimal given that the samples were only 1,000 ms in length). Each note was recorded into a Boss RC-3 Loop Station, which turns the electronic information from the keyboard into a digital audio file that could then be further processed. Each note was isolated and trimmed to 1,000 ms in Audacity, an open-source digital audio workstation. For the French Horn samples, because the patch did not include a natural decay, a fade-out effect was added just prior to the end of the audio sample to prevent clipping artifacts during playback. Each audio sample was exported (44.1 kHz, 16-bit) in a high-resolution lossless format (.wav files) to ensure sound quality. Although lossless audio files are larger in size than compressed formats (e.g., .mp3 and .ogg), all sounds were preloaded by the

experimental program to eliminate delays or glitches in playback related to file loading.

## Procedure

The procedure is divided into four subsections (see Fig. 1). The following sections describe each subsection of the procedure in greater detail.

### Study introduction

Upon initiating the study, each participant was randomly assigned to a training condition (diatonic, non-diatonic, mixed). The three training conditions were designed to test whether any differences in learning the target note (C4) could be attributed to the tonal context of training. The *diatonic condition* used non-target notes from the major scales of C, F, and G major, falling between F3 and B4. All three of these keys include C in their diatonic scales (as the tonic, dominant, and subdominant, respectively), meaning that the target note is congruent (i.e., has a high goodness-of-fit) in relation to the non-target notes (cf. Krumhansl, 1979). If C major was the selected context for the trial, the non-target notes were D4, E4, F4, G4, A4, and B4. If F major was the selected context for the trial, the non-target notes were F3, G3, A3, B*b*3, D4, and E4. If G major was the selected context for the trial, the non-target notes were G3, A3, B3, D4, E4, and F#4. One advantage of this approach is that the target note (C4) was not always in the same position relative to the non-target notes, and thus participants could not use an alternative strategy (e.g., simply listening for the lowest presented note) to recognize the target note. The *non-diatonic condition* used non-target notes from E, B, and F-sharp major scales falling between E3 and A#4. None of these keys include C diatonically, and thus the target note is incongruent (i.e., has a low goodness-of-fit) in relation to the non-target notes (cf. Krumhansl, 1979). If E major was the selected context for the trial, the non-target notes were E3, F#3, G#3, A3, B3, C#4, and D#4. If B major was the selected context for the trial, the non-target notes were B3, C#4, D#4, E4, F#4, G#4, and A#4. If F# major was the selected context for the trial, the non-target notes were F#3, G#3, A#3, B3, C#4, D#4, and E#4. Similar to the *diatonic* condition, this

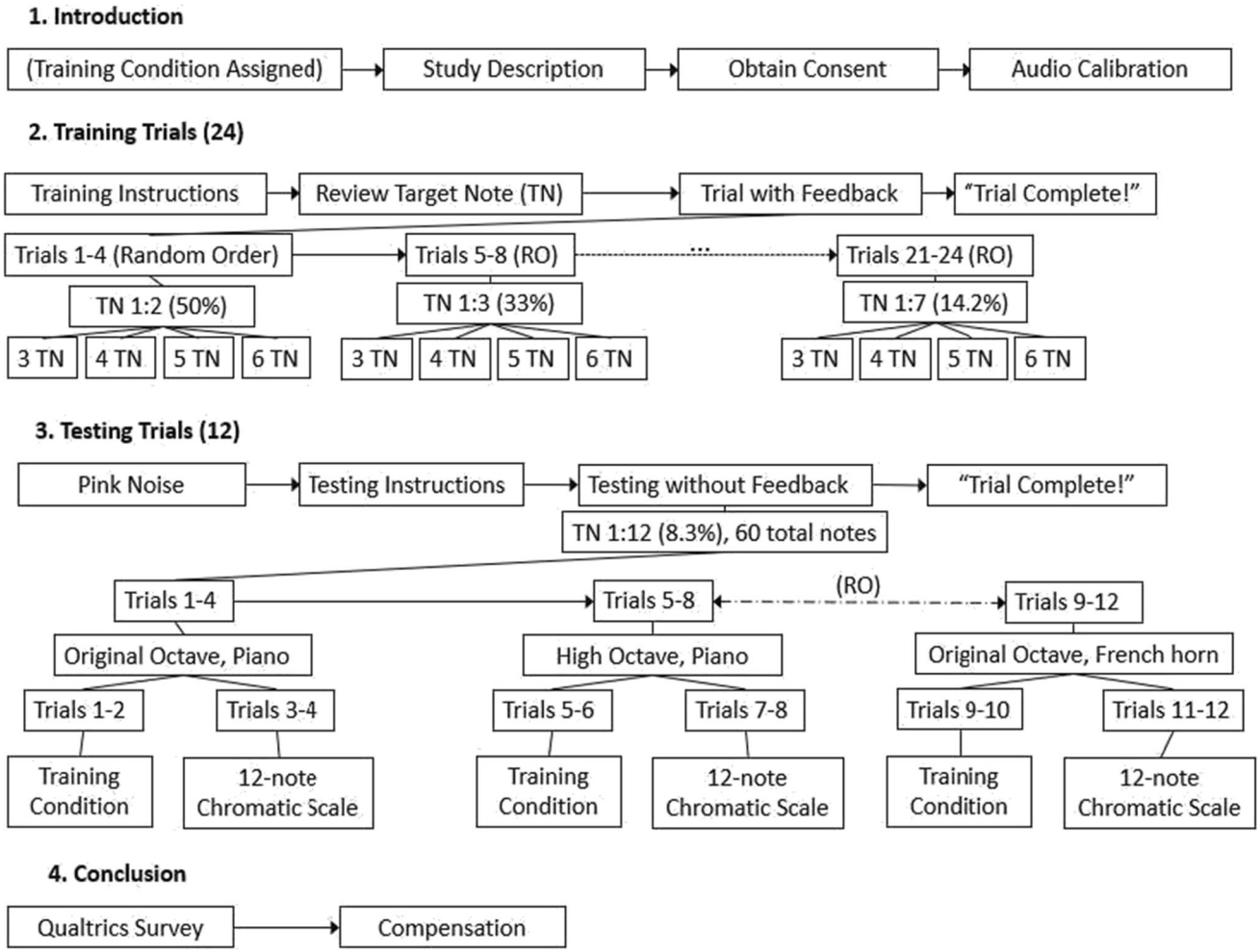


**Fig. 1** Overview of the procedure

approach allowed the target note to be heard in different positions relative to the non-target notes, and thus did not allow listening for the lowest or highest note to be an effective strategy. Finally, the *mixed condition* exposed participants to both the *diatonic* and the *non-diatonic* sets, with a particular scale context being randomly selected on each trial.

Participants were first presented with a digital form describing the study, and provided informed consent by clicking on a check box on the computer screen. Following the consent procedure, an auditory calibration exercise was used to (1) allow participants to adjust their computer volume to a comfortable listening level and (2) assess whether participants were following the experimenter's recommendations to wear headphones. The volume adjustment used a pink noise sample, RMS normalized to the same level as the musical notes, and participants pressed a button to continue with the study once their volume had been adjusted to a comfortable listening level. The headphone assessment was based on Woods et al. (2017) and consisted of six trials. On each trial, participants determined which of three tones sounded the quietest. The task is designed to be easy when participants have clear separation of left and right audio channels but virtually impossible if performed over internal computer speakers. Similar to Woods et al. (2017), we used a threshold of 83.33% (i.e., five of six correct answers) to determine headphone use. Given that headphone use was encouraged but not required, participants who failed this assessment ($n = 52$) were still included in the primary data analyses. Preliminary analyses including headphone use as a factor did not suggest that participants who used headphones differed in AP performance from participants who did not use headphones.

### Training trials

Next, participants completed the AP training portion of the experiment. The instructions stated that participants would be listening for a target note ("C"), which would be interspersed with non-target notes ("Not C"). Participants were instructed to press "C" or "N" on the keyboard if they thought the current note was a target note or non-target note, respectively. Feedback provided after each note notified participants whether the note was a target or non-target, and the font color of this feedback notified participants as to whether they had responded correctly (green) or incorrectly (red). For example, if participants saw "C" appear on the screen in a green font after a response, they would know that (1) the prior note was a target, and (2) they had correctly identified it as a target. If participants did not respond to the note within 1,500 ms, the script moved on automatically and the answer was marked as incorrect.

Before beginning each trial, participants were given an unlimited number of opportunities to hear the target note by clicking a button. Doing so was optional, and the number of times the subject opted to listen to the target note was recorded. The 24 training trials were binned into six difficulty levels of four trials each, with the training becoming more progressively difficult. For all conditions, each trial contained a different key from within the participant's assigned context in random, rotating order, such that each key was presented the same number of times by the end of training. For example, for the *congruent* condition, if the first trial was in C major, and the second was in G major, then the third would be in F major; this order would then repeat seven more times irrespective of difficulty level until the end of training.

In the initial difficulty level trials, the target note was presented 50% of the time. With each subsequent difficulty level, the relative proportion of the target note was reduced. The second set presented the target note 33% of the time, the third 25% of the time, until the final set, which presented the target note at 14.2%, or one in seven notes, equivalent to a random distribution among a diatonic scale. The four trials within each difficulty level contained different numbers of target notes: one trial contained three targets, one trial contained four targets, one trial contained five targets, and one trial contained six targets. Trial ordering was randomized within each difficulty level. The reason for varying the number of target notes across trials is because we did not want participants to develop an expectation that there was always the same number of targets (e.g., three), as this could encourage lapses of attention once participants realized that they had heard all the targets in a particular trial.

Within each trial, each note sounded for 1,000 ms (including decay time) and participants had 1,500 ms to make their response. Following this 1,500 ms response window, participants were briefly flashed feedback ("C" or "Not C", printed in green or red font) for 500 ms. After each trial, subjects were shown a "Trial Complete!" screen, which presented the opportunity to pause if desired. Across all training trials, participants responded to 108 target notes (18 per difficulty level) and 378 non-target notes (ranging from 18 in the first difficulty level to 108 in the final difficulty level).

### Testing trials

After completing the training trials, the subjects were tested on both specific and generalization learning. Participants listened to a 5,000-ms sample of pink noise prior to the beginning of each testing trial to discourage the use of echoic memory in the testing trials (cf. Darwin et al., 1972). The first four trials (*specific test*) used the same audio samples with which participants received training – piano notes in the same range as training. Participants were then administered two generalization tests, with the order randomly determined. The *octave generalization test* used piano notes that were shifted up one octave from the training context (making the target note C5). The *timbre generalization test* used the same range as the

training context but changed the instrument to the French horn. The first two trials of each subtest (specific, octave, timbre) used the same non-target notes from participants' training conditions – diatonic, non-diatonic, or mixed, and thus created a clear tonal context. The final two trials of each subtest drew non-target notes from all 11 non-C notes, thus disrupting any sense of key signature and providing a more stringent test of AP learning by using a chromatic context. In total, there were thus six types of testing trials (specific/tonal, specific/chromatic, octave/tonal, octave/chromatic, timbre/tonal, timbre/chromatic). Each testing trial contained five target notes and 55 non-target notes, meaning that the target note was not prioritized relative to the other non-target pitch classes (as the target note was heard at a 1:12 ratio). Thus, each of the six testing contexts contained ten target notes and 110 non-target notes.

Unlike training, no feedback was provided during testing. Participants had 1,500 ms to make their responses. Non-responses were marked as incorrect. After each test trial, subjects were shown a "Trial Complete!" screen and given an opportunity to take a break.

### Study session completion

Upon completion of the testing trials, participants were automatically redirected to a webpage to complete a Qualtrics survey containing questions on demographics and musical experience. Demographic questions included age, gender, and ethnicity. Musical experience questions included (1) whether participants played an instrument or sang, (2) the number of years of experience playing each reported instrument, collected in terms of six categories (*zero years, less than 1 year, between 1 and 2 years, between 2 and 5 years, between 5 and 10 years, and more than 10 years*), and (3) the number of hours per week spent playing each reported instrument, also collected in terms of six categories (*not actively playing, less than 1 h per week, between 1 and 5 h per week, between 5 and 10 h per week, between 10 and 15 h per week, more than 15 h per week*). Completion of the survey generated a unique code, which participants entered into Mechanical Turk and was manually verified by the experimenters to release payment.

## Data analysis

Performance in AP training and testing was operationalized using signal detection theory (Macmillan & Creelman, 2004). Specifically, using the "psycho" package in R (Makowski, 2018), we used d-prime (d') – a measure of sensitivity – to assess the extent to which participants were able to differentiate target notes from non-target notes. In the present context, a d-prime of 0 would represent an inability to differentiate "C" from other notes, whereas values above 0 would represent accurate discriminability and values below 0 would represent inaccurate discriminability (i.e., systematically labeling non-C notes as "C"). We additionally used the "psycho" package to calculate a measure of bias (c). The response bias measure of c is intuitive to interpret, with a value of 0 representing an "ideal observer" (i.e., equally weighting misses and false alarms). Negative values are reflective of a liberal threshold (i.e., the participant would respond "C" more often than the ideal observer), and positive values are reflective of a conservative threshold (i.e., the participant would respond "Not C" more often than the ideal observer). The values of c reflect the number of standard deviations from the position of the ideal observer. We included measures of both perceptual sensitivity (d') and response bias (c) because previous work has found that participants adopt more conservative criteria as a function of both increased task difficulty (lower d') and fatigue (Wylie et al., 2021). As such, we expected participants to become more conservative in their responses for the timbre and octave generalization tests in particular, as these were hypothesized to also have the lowest d-prime values.

We first assessed whether overall performance was above chance by comparing d' to 0 using one-sample *t*-tests. Separate d' values were calculated for each of the six difficulty levels of training, as well as for the three tests (specific, octave, timbre).

To assess the effects of condition on both training and testing performance, we used linear mixed-effects models via the "lme4" package (Bates et al., 2015). Separate models were constructed for training and testing performance, for both signal detection measures (d' and c). The training model contained difficulty level (1–6), condition (diatonic, non-diatonic, mixed), as well as the interaction of difficulty level and condition as fixed effects. The random-effects structure included participant intercepts, as well as slopes for difficulty level within each participant. The testing model contained condition (diatonic, non-diatonic, mixed), testing context (tonal, chromatic), and test type (specific, octave, timbre), as well as the interaction of these factors, as fixed effects. The random-effects structure included participant intercepts, as well as slopes for testing context and test type.

For both the training and testing models, we used the "MuMIN" package (Barton, 2020) to select the best fitting training and testing model. The best fitting model was a nested version of the global model and was selected based on the corrected Akaike Information Criterion (AICc; Akaike, 1998), which penalizes extra fitted parameters. As such, the selected model can be thought to represent the most parsimonious explanation of the current training and testing data.

To further explore participants' representations of the target note, we additionally examined the distribution of notes participants judged as "C" in terms of distance from C (in semitones). For example, a correctly judged "C" would yield a distance of 0, a "C#" judged to be the target note would yield a distance of +1, and a "B" judged to be

the target note would yield a distance of -1. We preserved directionality of errors, unlike prior investigations of AP using *absolute* deviation, collapsed across pitch class (e.g., Van Hedger & Nusbaum, 2018). This was motivated by our current research question (specifically relating to timbre and octave generalization), which requires a preservation of the directionality of responses in relation to the target note in order to assess the extent to which participants might be using pitch height as a cue.

Participant distributions of notes judged to be C were analyzed in two ways. First, we used bootstrapping, via the "boot" package (Canty & Ripley, 2021). For each bootstrap, we ran 1,000 simulations, in which we randomly sampled 1,000 responses (with replacement) to generate distributions of both the mean and the standard deviation of responses for each test. Significance was interpreted in terms of whether the averaged statistic for a given test (mean or standard deviation) fell within or outside of the 95% confidence intervals of the other tests. In this sense, the bootstrapping results help address whether participant distributions (1) were centered in the same place in terms of semitone distance, and (2) deviated from the target note with comparable magnitudes.

Second, we used the "mixtools" package in R (Benaglia et al., 2009) to assess the way in which each distribution could be understood in terms of finite mixture models. The present approach used an expectation-maximization algorithm to fit mixture models to the distributions from each test. The mixture models provided an understanding of the extent to which each distribution could be explained via two Gaussian distributions. This is particularly relevant in describing whether participants were systematically misclassifying a specific non-C note as the target note, which would result in more evenly weighted Gaussian distributions providing the best fit of the data. In contrast, if a distribution could be best explained in terms of a single Gaussian distribution (e.g., centered around the target note), then the model would heavily favor one Gaussian distribution over the other.

Based on the results of the bootstrapping and mixture models, which found that listeners' responses in the octave generalization task were bimodal with distributions centered both on the correct note (C5) as well as a perfect fourth below the correct note (G4, which was misclassified as the target note more often than the actual target note), we performed an exploratory analysis that examined whether misclassifying G4 as the target note varied across conditions for the octave generalization test. Responses in which G4 was labeled as the target were coded with a 1, and all other responses were coded with 0. This exploratory analysis only considered the chromatic trials; otherwise, the relative number of G4 notes would not be equated across the conditions. This exploratory analysis used a generalized linear mixed effects model with a binomial link. Participants were modeled with random intercepts.

Finally, to assess how musical experience and demographic variables related to training and testing performance, we used participants' mean d' values for each training and testing block and calculated Pearson correlation coefficients with (1) age, (2) musical training (yes/no), (3) whether one had experience playing the piano (yes/no), (4) number of years of experience playing one's primary instrument, and (5) whether one actively played an instrument (yes/no). Given the number of reported correlations, we used the Benjamini-Hochberg False Discovery Rate (FDR) alpha correction (Benjamini & Hochberg, 1995).

## Results

### Assessing performance against chance

All difficulty levels of training showed strong evidence of participants' abilities to accurately differentiate targets from non-targets (Table 1, top). This is perhaps not surprising given the fact that (1) participants were allowed to play the target note prior to each trial in training, and (2) feedback was provided after every note in training. For the test trials (Table 1, bottom), participants were robustly above chance for the specific test.

**Table 1** Summary of performance relative to chance estimates

| Assessment | Block | Target % | d' | 95% CI | *t* | Cohen's d |
|---|---|---|---|---|---|---|
| *Training* | Level 1 | 50.0% | 2.81 | [2.67, 2.96] | 39.09*** | 2.94 |
| | Level 2 | 33.3% | 2.99 | [2.84, 3.14] | 39.71*** | 3.01 |
| | Level 3 | 25.0% | 2.71 | [2.54, 2.87] | 32.65*** | 2.48 |
| | Level 4 | 20.0% | 2.50 | [2.34, 2.67] | 29.70*** | 2.25 |
| | Level 5 | 16.7% | 2.33 | [2.16, 2.51] | 26.29*** | 1.99 |
| | Level 6 | 14.3% | 2.25 | [2.09, 2.42] | 27.00*** | 2.05 |
| *Testing* | Specific | 8.33% | 1.24 | [1.08, 1.41] | 14.87*** | 1.12 |
| | Timbre | 8.33% | 0.58 | [0.44, 0.72] | 8.33*** | 0.64 |
| | Octave | 8.33% | 0.23 | [0.09, 0.37] | 3.34** | 0.25 |

*Note*: Target % refers to the relative percentage of target notes relative to non-target notes. *** $p < .001$, ** $p < .01$

Performance on the timbre generalization test was attenuated relative to the specific test but was still significantly above chance. Performance on the octave generalization test was attenuated relative to both the specific and timbre generalization tests but was still above chance. Performance on each test is further explored in the next two subsections.

## Modeling performance as a function of condition and test type

### Discriminability (d')

The model for the training data (Fig. 2A) showed a significant negative effect of difficulty level (B = -0.08, *SE* = 0.024, $p < .001$), with performance decreasing as a function of increased difficulty level. Relative to the diatonic condition, which was treated as the reference condition, participants in both the non-diatonic and mixed conditions performed comparably overall ($ps > .124$). However, condition interacted with difficulty level, with participants in both the non-diatonic (B = -0.11, *SE* = 0.032, $p < .001$) and mixed (B = -0.08, *SE* = 0.034, $p = .018$) conditions showing significantly worsening discriminability as a function of difficulty compared to the diatonic participants. The approach of selecting the best-fitting nested model from this training model resulted in a model that only retained difficulty level, which was highly significant (B = -0.15, *SE* = 0.014, $p < .001$). Thus, despite the significant effects of condition reported in the primary model, these terms did not outweigh the penalty for adding extra parameters to the model.

The model for the testing data (Fig. 2B) showed that, relative to the specific test, performance on both the timbre generalization (B = -0.88, *SE* = 0.15, $p < .001$) and octave generalization (B = -1.12, *SE* = 0.16, $p < .001$) tests was significantly worse. A Tukey-corrected post hoc test using the "emmeans" package (Lenth, 2021) additionally demonstrated that performance in the timbre generalization test was significantly higher than performance in the octave generalization test (B = 0.35, *SE* = 0.06, $p < .001$). Thus, all three tests were differentiated from one another in terms of performance. In terms of condition, participants in the mixed condition performed overall worse than participants in the diatonic condition (B = -0.49, *SE* = 0.22, $p = .028$). Performance was

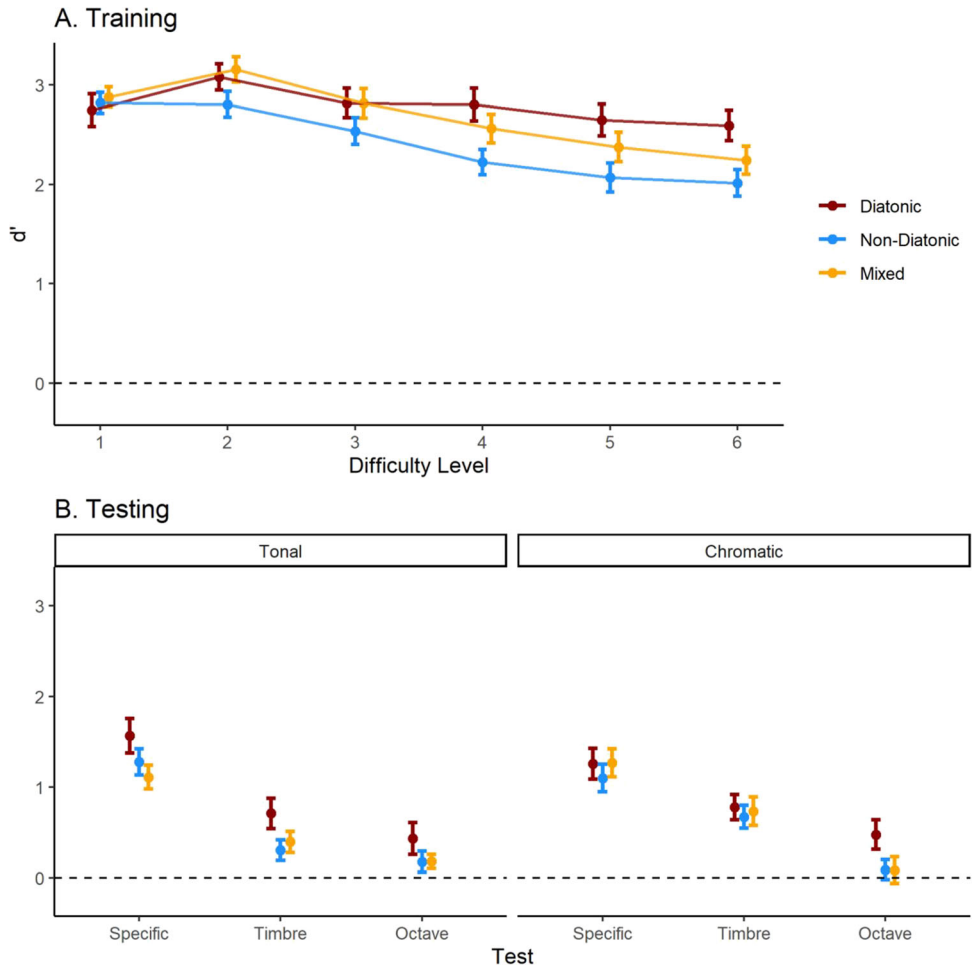


**Fig. 2** Summary of mean training and testing performance (d-prime) as a function of condition. Error bars represent ±1 standard error of the mean. The dashed line represents chance performance

additionally overall worse for trials that used a chromatic context compared to trials that used a tonal context (B = -0.31, *SE* = 0.12, *p* = .011).

There was a significant interaction between testing context and the mixed training condition relative to the diatonic training condition (B = 0.50, *SE* = 0.17, *p* = .003). This interaction was characterized by worse performance for diatonic participants in the chromatic (d' = 0.84) versus tonal (d' = 0.92) contexts, as opposed to *better* performance for mixed participants for the chromatic context (d' = 0.70) relative to the tonal context (d' = 0.59). Chromatic context additionally interacted with both the timbre generalization (B = 0.40, *SE* = 0.16, *p* = .013) and the octave generalization (B = 0.33, *SE* = 0.16, *p* = .038) tests, relative to the specific test. These interactions were characterized by attenuated performance in the specific test when listening for the target note in a random (d' = 1.20) compared to a tonal (d' = 1.25) context, as opposed to better performance in the timbre generalization test when listening for the target note in a random (d' = 0.72) compared to a tonal (d' = 0.46) context. In the octave generalization test, performance was more attenuated when listening for the target note in a random (d' = 0.21) compared to a tonal (d' = 0.27) context. Finally, there was a three-way interaction between condition (mixed relative to diatonic), chromatic context, and test (octave generalization relative to specific). This interaction can be broken down as follows: for the diatonic participants, a random tonal context impaired performance relative to a tonal context in the specific test (d' = 1.25 and 1.57, respectively), whereas a random tonal context nominally facilitated performance relative to a tonal context in the octave generalization test (d' = 0.47 and 0.43, respectively). In contrast, the mixed participants showed an entirely opposite pattern of results. A random tonal context *facilitated* performance relative to a tonal context in the specific test (d' = 1.27 and 1.04, respectively), whereas a random tonal context impaired performance relative to a tonal context in the octave generalization test (d' = 0.08 and 0.24, respectively). No other term was significant in the model. The best-fitting nested model contained test type (specific, timbre generalization, octave generalization), testing context (chromatic, tonal), and the interactions of these factors. In the best-fitting model, the main effects of the timbre generalization and octave generalization tests (relative to the specific test), as well as the interaction of testing context and the timbre generalization test (relative to the specific test) were significant. The main effect of testing context, as well as the interaction of the octave generalization test (relative to the specific test) were not significant.

### Response bias (c)

We predicted that response bias would become more conservative as task difficulty and time on task increased (cf. Wylie et al., 2021). This prediction was supported in the present data. The model for the training data (Fig. 3A) showed a significant positive effect of difficulty level (B = 0.075, *SE* = 0.011, *p* < .001), with response bias showing a more conservative trend as a function of difficulty level. No other term was significant in the model (all *p*s > .235). The approach of selecting the best-fitting nested model from this training model resulted in a model that only retained difficulty level, which was highly significant (B = 0.08, *SE* = 0.006, *p* < .001).

The model for the testing data (Fig. 3B) showed that, relative to the specific test, response bias on both the timbre generalization (B = 0.20, *SE* = 0.07, *p* = .008) and octave generalization (B = 0.29, *SE* = 0.07, *p* < .001) tests was significantly more conservative. A Tukey-corrected post hoc test using the "emmeans" package (Lenth, 2021) additionally demonstrated that response bias in the octave generalization test was significantly more conservative than response bias in the timbre generalization test (B = 0.17, *SE* = 0.03, *p* < .001). Thus, similar to the sensitivity analyses, all three tests were differentiated from one another in terms of response bias (with bias becoming increasingly conservative going from specific to timbre generalization to octave generalization tests). Participants in the mixed condition were also more conservative in their responses compared to participants in the diatonic condition (B = 0.30, *SE* = 0.07, *p* < .001). In terms of higher-order interactions, similar to the sensitivity analyses, there was a significant interaction between testing context and condition (mixed vs. diatonic; B = -0.42, *SE* = 0.08, *p* < .001). This interaction was characterized by a more conservative approach for diatonic participants going from a tonal to a chromatic context (c = 0.88 and 0.94, respectively). In contrast, participants in the mixed condition became relatively more liberal going from a tonal to a chromatic context (c = 1.19 and 0.87, respectively). There was also a significant two-way interaction between condition (non-diatonic vs. diatonic) and test (octave vs. specific; B = 0.20, *SE* = 0.09, *p* = .031). This interaction was characterized by both conditions becoming more conservative for the octave generalization test relative to the specific test, with the non-diatonic participants becoming even more conservative (c = 0.78 to 1.15) compared to the diatonic participants (c = 0.76 to 1.05). There were additionally two three-way interactions that reached significance. First, there was a three-way interaction between condition (non-diatonic vs. diatonic), testing context, and test (timbre generalization vs. specific; B = -0.30, *SE* = 0.11, *p* = .005). This interaction was characterized by both conditions becoming more conservative when tested in a tonal context going from the specific test to the timbre generalization test (c = 0.73 to 0.91 for the diatonic participants and c = 0.70 to 1.05 for the non-diatonic participants). In contrast, when tested in a chromatic context, diatonic participants became more conservative going from the specific to the timbre generalization test (c = 0.81 to 0.95), whereas non-diatonic participants were relatively stable in their response bias (c = 0.86 to 0.89). Second, there was a three-way interaction between condition (non-

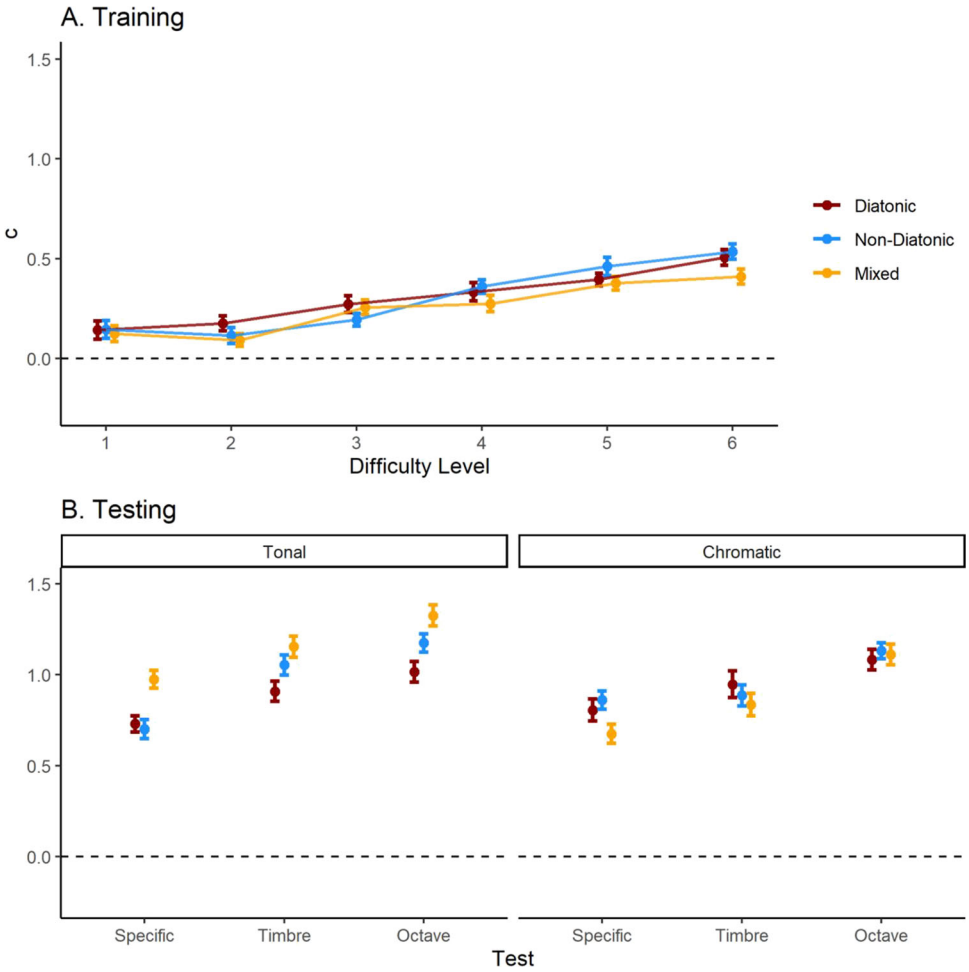


**Fig. 3** Summary of mean training and testing response bias (c) as a function of condition. Error bars represent ±1 standard error of the mean. The dashed line represents an ideal observer (i.e., identical proportions of misses and false alarms). Values above the dashed line reflect a conservative bias (i.e., having a higher proportion of misses compared to false alarms)

diatonic vs. diatonic), testing context, and test (octave generalization vs. specific; B = -0.23, *SE* = 0.11, *p* = .034). This interaction was characterized by both conditions becoming comparably more conservative when tested in a chromatic context going from the specific test to the octave generalization test (c = 0.81 to 1.08 for the diatonic participants and c = 0.86 to 1.13 for the non-diatonic participants). In contrast, when tested in a tonal context, non-diatonic participants became much more conservative going from the specific to the octave generalization test (c = 0.70 to 1.17) compared to diatonic participants (c = 0.73 to 1.02).

No other term was significant in the model. The best-fitting nested model contained test type (specific, timbre generalization, octave generalization), testing context (chromatic, tonal), condition (diatonic, non-diatonic, mixed), and the two-way interaction between testing context and condition. In the best-fitting model, the main effects of the timbre generalization and octave generalization tests (relative to the specific test), as well as the interaction of testing context and the mixed condition (relative to the diatonic condition) were significant.

## Analysis of response distributions

Although signal detection measures provide an assessment of sensitivity and response bias, they do not inherently provide a detailed examination of *how* participants incorrectly classified notes (as all "false alarms" are treated as equivalent). To better characterize the distributions of responses in which participants reported hearing the target note, we therefore analyzed each response in relation to the target note. We began by simply plotting the descriptive statistics (histograms of deviations from the target note) in Fig. 4A. These histograms reveal a striking difference between the octave generalization test and both the specific and the timbre generalization tests. Unlike the specific and timbre generalization distributions, which appear unimodal and centered on 0 (i.e., the target note), the octave generalization distribution is skewed right, with a modal response of -5 semitones (i.e., G4). This qualitative assessment of the distributions is quantitatively explored in the subsequent paragraphs of this subsection.

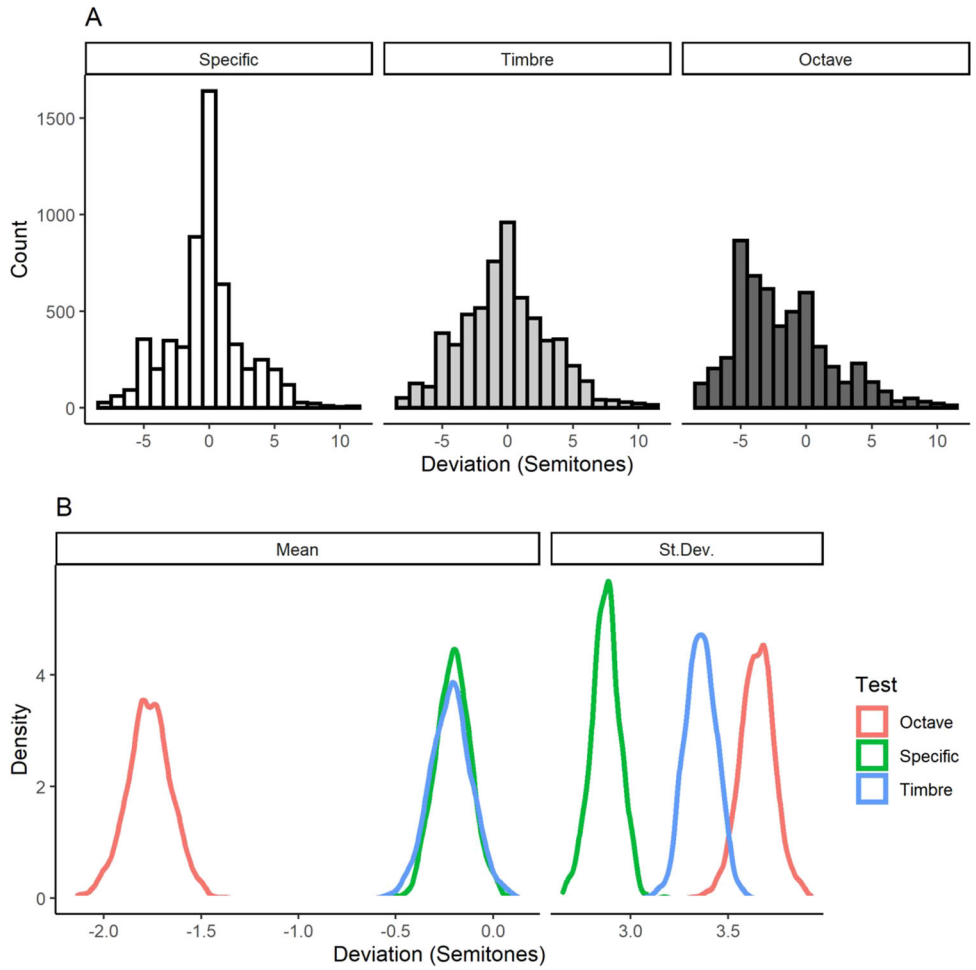


**Fig. 4** Histograms of performance in relation to the target note. Panel A depicts individual responses as a function of Test. Panel B depicts the results from the bootstrapped means (left) and standard deviations (right), with Test represented by color

The bootstrapped mean from the specific test was -0.19 semitones (95% CI: [-0.35, 0.01]). The bootstrapped mean from the timbre generalization test was -0.08 semitones (95% CI: [-0.15, 0.27]), which fell within the confidence interval of the specific test. In contrast, the bootstrapped mean for the octave generalization test was -1.88 semitones (95% CI: [-2.21, -1.76]). The octave generalization test was the only distribution not including zero in its confidence interval, and its confidence interval was non-overlapping with both the specific and timbre test distributions.

The bootstrapped standard deviation from the specific test was 2.83 semitones (95% CI: [2.64, 2.93]). The bootstrapped standard deviation from the timbre generalization test was 3.43 semitones (95% CI: [3.34, 3.66]), which fell outside of the 95% confidence interval of the specific test and thus suggests greater variability in responses. The bootstrapped standard deviation from the octave generalization test was 3.76 semitones (95% CI: [3.71, 4.05]). Taken together, the results from the bootstrapping can be summarized in the following way. First, relative to the specific test, responses in the timbre generalization test were distributed similarly around the target note but were significantly more variable. Second, responses in the octave generalization test were significantly shifted towards the lower end of the distribution and were significantly more variable than both the specific and timbre generalization tests. The results from the bootstrapping analyses are plotted in Fig. 4B.

The mixture model fit to the specific test suggested that the data were somewhat well characterized by two Gaussian distributions, with the first distribution contributing 32.2% to explaining the specific test data and the second distribution contributing 67.8%. However, these two distributions were not differentiated in terms of their mean value (-0.13 vs. -0.24 semitones, respectively); rather the distributions were differentiated in terms of their variance. The first distribution had relatively little variance ($SD$ = 0.59), whereas the second distribution had a much wider variance ($SD$ = 3.46). The two Gaussian distributions are overlaid on the specific test data in Fig. 5A. The mixture model fit to the timbre generalization test suggested that the data were not as well characterized by two Gaussian distributions, with the model failing to converge. The first distribution ($M$ = -0.91, $SD$ = 2.97)

contributed 78.9% to explaining the timbre generalization test data. The second distribution ($M = 2.40$, $SD = 3.47$) only contributed 21.1% to explaining the data. The two Gaussian distributions are overlaid on the timbre generalization test data in Fig. 5B. The mixture model fit to the octave generalization test suggested that the data were well explained by two Gaussian distributions. The first distribution explained 40.0% of the octave generalization test data and was several semitones lower than the target note ($M = -4.37$, $SD = 1.60$). The second distribution explained 60.0% of the data and was centered around the target note ($M = -0.03$, $SD = 3.60$). The two Gaussian distributions are overlaid on the octave generalization test data in Fig. 5C.

Based on the finding that participants systematically misclassified G4 as C5 in the octave generalization test, we conducted an exploratory analysis to assess whether this result differed across training conditions. Results from this analysis suggested that the odds of misclassifying G4 as C5 varied as a function of condition. Specifically, compared to the diatonic training condition (which was used as the reference category), participants in the non-diatonic condition (B = -0.73, $SE$ = 0.26, $p$ = .005) had lower log odds of misclassifying G4 as C5, despite limiting the analyses to chromatic test trials in which both groups were equally likely to hear G4. In contrast, the mixed condition (B = -0.22, $SE$ = 0.27, $p$ < .410) did not differ from the diatonic condition with respect to misclassifying G4 as C5. A post hoc comparison between the non-diatonic and mixed training conditions showed that the non-diatonic condition had marginally lower odds of misclassifying G4 as C5 (B = -0.51, $SE$ = 0.26, $p$ = .073).

### Musical experience and performance

The results from the exploratory correlational analyses are reported in Table 2. The musical experience measures were positively intercorrelated, as were the d' values from several of the training blocks. Performance on the specific test was also significantly correlated with performance on training blocks 2–6. However, performance on the timbre and octave generalization tests were not significantly correlated with either performance on the specific test or with performance on any of the training blocks. Additionally, we did not find any evidence that either the demographic or the musical experience measures related to training or testing performance.

## Discussion

Adult participants clearly learned to differentiate "middle C" (C4) from other notes, regardless of the tonal context used in training (see Fig. 2B). Furthermore, in testing, participants were able to recognize C4 outside of *any* tonal context (i.e., the chromatic test trials), which represents a more stringent test of AP by presenting all 12 pitch classes with equal probability. Participants were also able to partly generalize to an unheard instrument, as evidenced by the above-chance (albeit attenuated) performance in the test using a French horn timbre that was not presented during training. This finding suggests that the skill acquired in the present study would not be considered instrument-specific AP (cf. Reymore & Hansen, 2020).

However, the most striking finding of the present study is that the current AP training paradigm resulted in relatively weak generalization across octaves, even relative to generalization across timbres (see Fig. 2B). Although sensitivity was above chance for octave generalization, the d-prime was close to zero and the effect size was small. This finding has several immediate implications; notably, it demonstrates the importance of including testing across octaves as a necessary element of any valid AP test. The fact that participants were able

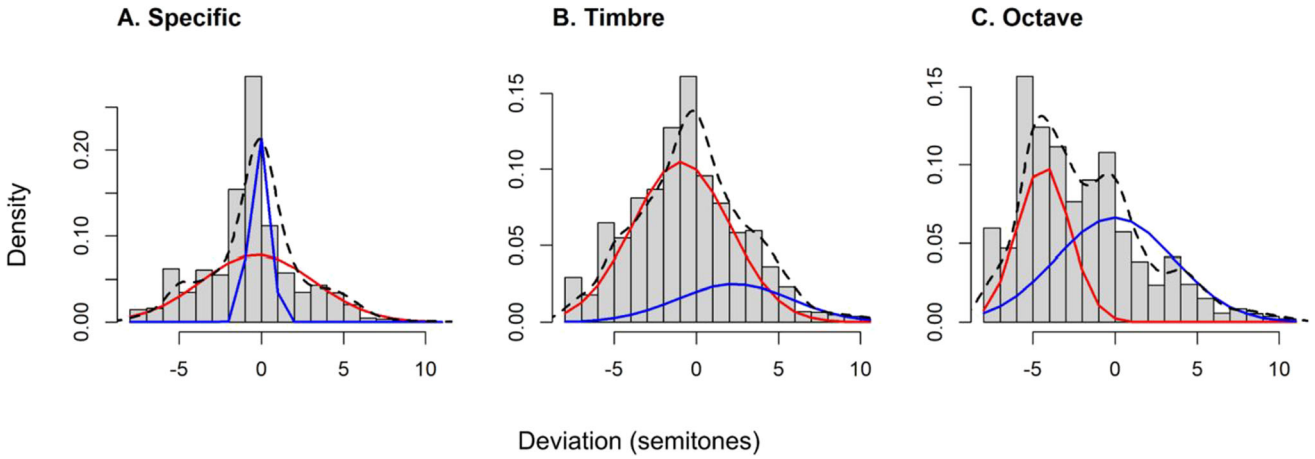


**Fig. 5** Gaussian mixture models applied to each test. The solid red and blue lines outline the two Gaussian distributions that provided the best fit of the histograms for each test. The dotted black line outlines the overall shape of the distribution. A deviation of 0 corresponds to correctly identifying a note as the target note

**Table 2** Correlation matrix of questionnaire measures and performance (d')

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. Age | - | | | | | | | | | | | | | |
| 2.Musical Training | -.06 | - | | | | | | | | | | | | |
| 3.Piano | .04 | **.62** | - | | | | | | | | | | | |
| 4.Years of Training | .06 | **.92** | **.61** | - | | | | | | | | | | |
| 5.Active Musician | .01 | **.72** | **.54** | **.75** | - | | | | | | | | | |
| 6.Training - Block 1 | -.03 | .03 | .06 | .11 | .05 | - | | | | | | | | |
| 7.Training - Block 2 | -.01 | .36 | .36 | .44 | .39 | .50 | - | | | | | | | |
| 8.Training - Block 3 | .06 | .20 | .20 | .27 | .16 | **.55** | **.72** | - | | | | | | |
| 9.Training - Block 4 | .12 | .39 | .33 | .49 | .32 | .38 | **.72** | **.69** | - | | | | | |
| 10.Training - Block 5 | .04 | .18 | .20 | .26 | .14 | .37 | **.71** | **.74** | **.75** | - | | | | |
| 11.Training - Block 6 | .03 | .25 | .29 | .35 | .15 | .41 | **.72** | **.76** | **.76** | **.83** | - | | | |
| 12.Specific Test | .05 | .10 | .15 | .23 | .05 | .29 | **.61** | **.66** | **.71** | **.72** | **.79** | - | | |
| 13.Octave Test | .17 | .20 | .04 | .23 | .08 | .06 | .16 | .16 | .25 | .10 | .20 | .25 | - | |
| 14.Timbre Test | .10 | .22 | .23 | .24 | -.04 | .04 | .35 | .37 | .37 | .41 | .49 | .43 | .44 | - |

*Note:* Significant correlations, using an FDR correction, are highlighted in bold. 1: Age of participant, 2: Reported musical training (1/0), 3: Reported piano training (1/0), 4: Years of training on primary instrument, 5: Currently active musician (1/0)

to demonstrate learning across a variety of tonal contexts and generalize across timbre but *not* octave to the same degree suggests that participants did not use pitch chroma exclusively, but rather some other learning strategy (such as pitch height, discussed in detail below). This weak generalization across octaves also implies that although both timbre and octave generalization are necessary components of "genuine" AP ability, they are likely achieved via different mechanisms or listening strategies.

A detailed analysis of the distribution of participant responses demonstrated that participants used pitch *height* and relative pitch alongside pitch *chroma* in recognizing C (see Fig. 4A). These strategies would result in high levels of performance when pitch height and pitch chroma were not dissociable (i.e., the specific and timbre generalization tests), but lower levels of performance when pitch height and pitch chroma were dissociable (i.e., the octave generalization test). In support of this conjecture, whereas the modal response in the specific and timbre generalization tests was centered around the target note C4, the modal response in the octave tests was a deviation of -5 semitones from the target note C5 (i.e., G4). The choice of one of the lowest pitches in the octave suggests that subjects were looking for a pitch closest to the range in which they were trained (i.e., the use of pitch height as a strategy). This pattern of results is also not easily explained by participants simply being confused about the task; the instructions prior to the octave generalization test explicitly stated that the Cs would sound higher than the ones heard in training, and the results of the Gaussian mixture models found evidence that some responses were correctly distributed around the target note of C5 (Fig. 4C). Given that pitch chroma and pitch height are dissociable constructs with differentiable neural pathways (Warren et al., 2003), it is possible that participants defaulted to using pitch height as a cue, which would not have provided any benefit to categorization based on chroma. The important question that stems from this finding is whether the task could be manipulated in a manner to emphasize pitch chroma over pitch height for octave generalization, or whether most listeners are simply unable to attend to pitch chroma in listening contexts when pitch height is varied. For example, future work might consider strategies for framing octave generalization to facilitate understanding the task in terms of pitch chroma (e.g., to engage in explicit imagery processes of imagining what a higher or lower C would sound like). If performance improves considerably with this kind of intervention, it suggests that participants in the present experiment might be able to categorize based on pitch chroma, even in a novel pitch height range. In contrast, if this kind of intervention does not improve octave generalization performance, this might suggest that genuine AP training may only be possible for some individuals (e.g., see Van Hedger et al., 2019).

Although the use of pitch height alone would have led participants to simply select the lowest pitch of the set (which was E4 for the octave generalization test), we suggest that a relative pitch strategy was also used due to the fact that G4 was the lowest pitch of the set that had a strong relative tonal relationship with the target note (e.g., see Krumhansl, 1979). G and C are separated by a perfect fifth, which share a privileged relationship in Western tonal music. In addition to participants rating pitch classes separated by perfect fifths to have high congruence/goodness-of-fit (e.g., Krumhansl, 1979), corpus analyses have found that the perfect fifth is the second most common note in tonal contexts apart from

the tonic (Temperley & de Clercq, 2013). Furthermore, novel song melodies that contain a higher proportion of perfect fourths and fifths are more likely to be falsely remembered as being in the original key of the song (Van Hedger et al., 2022), suggesting a role of the tonal hierarchy on the learning and memory of musical events. Thus, it is reasonable to conclude that listeners learned to attend to pitch height based on the training (which did not vary octave) and, when this cue was disrupted, relied on some combination of pitch height and relative pitch (e.g., selecting the pitch that was closest to C4 in both pitch height and psychological distance).

It is also interesting to note that the (mis)classification of G4 as C5 was augmented in the diatonic training condition, at least relative to the non-diatonic training condition. This finding further supports the idea that participants used tonal context in the present training paradigm, even if tonal context did not lead to overall changes in detecting the target note. On the one hand, it is perhaps not surprising that participants in the diatonic training condition mislabeled G4 as C5 more often than those in the non-diatonic condition, as the former participants would have heard G4 in all trained keys (C, G, F) during training given its close relationship to C. In fact, participants in the non-diatonic condition did not receive *any* experience with the pitch class G until chromatic testing, as the trained key signatures all did not contain G (E, F#, B). On the other hand, given that feedback was provided after every note in training, it is also reasonable to expect that participants in the diatonic condition would have been *less* susceptible to this misclassification, as they would have received more explicit feedback during training about pitch class G. In this sense, the heightened misclassification of G4 as C5 for the diatonic condition participants might represent a kind of "false memory" based on tonal similarity (cf. Vuvan et al., 2014).

Analyses of response bias also demonstrated systematic changes as a function of both training difficulty level and test. Generally, participants became more conservative (i.e., less likely to classify a note as "C") as difficulty increased. This pattern was apparent in both training and testing, with increasingly conservative response bias going from the specific test to timbre generalization to octave generalization. This overall pattern is perhaps not surprising, particularly for the timbre and octave generalization tests, as the targets in these tests were never experienced in training and thus participants may have been more hesitant to classify any sound as the target.

Despite the understanding in AP research that generalization across both octaves and timbres is a critical feature of "genuine" AP (e.g., Bachem, 1937; Reymore & Hansen, 2020), the explicit testing of both of these dimensions is inconsistent in the literature. Significantly, the closest analogues of the current training paradigm fail to test octave or timbre generalization at all, and yet are interpreted as providing empirical evidence for "critical period" mechanisms of AP acquisition. Crozier (1997) trained groups of kindergarteners and ninth-graders to identify A4 (440 Hz) for 5 min a day, ultimately finding that the kindergarteners outperformed the ninth-graders. Testing involved reproducing A4 from memory, then identifying A4 from a set of three notes six times. However, non-target notes were always at least four semitones away from the target note in height. Crozier (1997) suggests that the distancing of the target note from the non-target notes calls into question whether participants were relying on either pitch height or relative pitch to discern the target note instead of pitch chroma. Russo et al. (2003) expanded upon Crozier (1997) by implementing several methodological changes – namely, training participants individually rather than as a part of a group, comparing younger (3- to 4-year-old) and older (5- to 6-year-old) children to adults, and testing the target note (C4) with six non-target notes, without a pitch gap between the non-target and target notes. Russo et al. (2003) conceptually replicated Crozier (1997) in finding that the 5- to 6-year-old participants outperformed both the 3- to 4-year-olds and the adults.

Although both of the above studies have been used to make claims with respect to mechanisms of AP acquisition, neither study assessed generalization beyond the exact trained note (either in terms of timbre or octave generalization) and thus, based on the present findings, cannot be used to make strong claims that pitch *chroma* was learned. Crozier (1997) argues in favor of the critical period theory in addition to other mechanisms for learning AP, but at the same time concedes that disentangling the contributions of relative and absolute pitch as potential strategies was not possible given the paradigm. Russo et al. (2003) claims the study "provide[s] strong support for a critical period for absolute pitch acquisition" (p.119), despite not testing how learning generalized across timbres or octaves. As such, the findings of the present experiment contextualize these prior investigations of AP training. Had an octave generalization test *not* been included in the present study, the conclusion that robust, adult AP learning may have been erroneously reached. The present experiment therefore emphasizes the need to both train and test for AP in a manner that is consistent with its definition, ensuring that participants are able to successfully demonstrate *all* the skills of a typical AP possessor before being labeled as such. Not doing so may result in measuring learning of a different dimension, such as pitch height, the strategy we believe the present results demonstrate the participants were using.

Although the argument that AP testing must incorporate timbre and octave generalization appears to be limited to the field of music perception and cognition, the present findings have broader implications for measurement in psychology and behavioral science. Any experimental investigation must operationalize a complex construct (e.g., executive functions, personality) in terms of quantifiable, performance metrics on a task. However, it is important to ensure that any given task is adequately capturing the presumed dimensions of interest for

a given construct in a theory-driven manner – i.e., to ensure construct validity of a given measurement (e.g., Smith, 2005). In the present context, it is clear that a theory-driven measure of AP would necessarily dissociate pitch height from pitch chroma, which is accomplished through testing multiple octaves. However, these specific findings reassert broader conclusions that psychological tasks need to be designed in a theory-driven manner and additionally should not be conflated with the construct of interest.

The present study does not inherently refute or support a critical period theory of AP acquisition. It does, however, suggest that octave and timbre generalization should be considered necessary components of any training study seeking to make claims about AP learning mechanisms, regardless of the age of participants. As the present results make clear, training and testing a single frequency may not even be complex enough for what could be considered "proof-of-concept" AP learning, as learning can manifest even if participants attend to the wrong cue; without testing multiple octaves, it is impossible to distinguish whether subjects are learning pitch height or pitch chroma. Future studies could expand upon the findings of Crozier (1997) and Russo et al. (2003), for example, by testing how children and adults perform at octave generalization tasks. If indeed, as both findings suggest, children are able to outperform adults in AP generalization tests, this would provide a more rigorous standard of evidence to support a critical period theory of AP acquisition.

The present study aimed to examine the difference in learning accuracy in various tonal contexts presented during training; however, the results suggested that tonal context had a small effect in training and no significant effect in testing. This lack of an effect in testing was unexpected; however, because the present findings suggest participants were attending to pitch height and not pitch chroma, it is possible that the present results are not well-suited to address the role of tonal context in AP learning. There was some evidence that training in a diatonic context resulted in improved learning, particularly at the most difficult levels of training (see Figure 2A). This finding could be argued in support of Brady's (1970) fixed-scale theory of AP learning; he suggests that some people may learn pitches from one scale, and then use their relative relationships to an anchor note (i.e., the tonic) to form chroma even when not using the same scale. Future work building upon the present study would be of value; particularly, an adaptation of the training program that includes target Cs in multiple octaves would eliminate the subjects' option to rely solely on pitch height, the results of which would be better suited to analyze the relationship between context and learning.

Only a minority of participants reported any prior musical training (see Fig. S1 in the Online Supplemental Material (OSM) for a histogram), making the observed learning more notable. Although the associations between musical training and AP learning in the present study were nominally positive (Table 2), they were relatively weak and did not survive FDR corrections. This lack of an association at first glance may appear surprising, yet it can be interpreted in a larger context of prior work associating musical training to pitch chroma representations among non-AP possessors. Van Hedger et al. (2015a) found that auditory working memory mediated the relationship between a musical measure (age of beginning musical training) and explicit AP learning. Furthermore, absolute pitch memory for popular recordings has similarly been demonstrated among nonmusicians (Schellenberg & Trehub, 2003), with a related experiment concluding that there is no significant association between explicit musical training and absolute pitch memory for familiar recordings (Van Hedger et al., 2018a). These findings may provide insight into the observed learning in this explicit AP training paradigm, which similarly did not seem to be strongly associated with aspects of formal musical training. However, it is also important to note that our musical questions might not have been sensitive enough to detect associations, as we only asked participants to describe their musical training within relatively broad categories (e.g., "5 to 10 years") and our sample was zero inflated (i.e., the majority of participants reported no musical training).

Overall, the present study found that a graded difficulty learning paradigm can rapidly teach participants to distinguish C4 from other notes, and that this learning is still demonstrable when the target note is presented with equal likelihood relative to any other note (i.e., independent of a tonal context) and when feedback is not provided. Participants were able to partly generalize across instruments, demonstrated by significantly attenuated (but still above-chance) performance relative to the trained instrument. Most importantly, the present study concludes that the participants' weakened ability to generalize across octaves suggests that both present and prior paradigms that train a single target frequency are inappropriate measures of AP learning – even as proof-of-concept demonstrations. Without an assessment of octave generalization, participants can rely on alternative strategies (e.g., attending to pitch height) that cannot be disentangled from pitch chroma. The present findings therefore argue in favor of obligatory testing that spans timbre and octave for any work that seeks to assess AP training and performance. Doing so will give clarity to the mechanisms that are actually being used in any attempts to train AP, which will strengthen the literature for and foster better understanding of the learning and maintenance of AP across the lifespan.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.3758/s13414-023-02653-0.

**Funding** This work was supported in part from an internal research grant (SVH).

## Declarations

**Conflicts of interest** The authors have no conflicts of interest to declare.

## References


Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences, 8*(10), 457–464. https://doi.org/10.1016/j.tics.2004.08.011

Akaike, H. (1998). Information Theory and an Extension of the Maximum Likelihood Principle. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), *Selected Papers of Hirotugu Akaike* (pp. 199–213). Springer. https://doi.org/10.1007/978-1-4612-1694-0_15

Bachem, A. (1937). Various types of absolute pitch. *The Journal of the Acoustical Society of America, 9*(2), 146–151. https://doi.org/10.1121/1.1915919

Bachem, A. (1955). Absolute pitch. *The Journal of the Acoustical Society of America, 27*(6), 1180–1185. https://doi.org/10.1121/1.1908155

Baharloo, S., Service, S. K., Risch, N., Gitschier, J., & Freimer, N. B. (2000). Familial Aggregation of Absolute Pitch. *The American Journal of Human Genetics, 67*(3), 755–758. https://doi.org/10.1086/303057

Barton, K. A. (2020). *MuMIn: Multi-model inference*. R package version 1.43.17. https://CRAN.Rproject.org/package=MuMIn

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Benaglia, T., Chauveau, D., Hunter, D. R., & Young, D. (2009). mixtools: An *R* Package for analyzing finite mixture models. *Journal of Statistical Software, 32*(6), 1–29. https://doi.org/10.18637/jss.v032.i06

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B: Methodological, 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Brady, P. T. (1970). Fixed-Scale Mechanism of Absolute Pitch. *The Journal of the Acoustical Society of America, 48*(4B), 883–887. https://doi.org/10.1121/1.1912227

Brammer, L. M. (1951). Sensory cues in pitch judgment. *Journal of Experimental Psychology, 41*(5), 336–340. https://doi.org/10.1037/h0058974

Brown, W. A., Sachs, H., Cammuso, K., & Folstein, S. E. (2002). Early Music Training and Absolute Pitch. *Music Perception, 19*(4), 595–597. https://doi.org/10.1525/mp.2002.19.4.595

Canty, A., & Ripley, B. D. (2021). *boot: Bootstrap R (S-Plus) Functions* (R package version 1.3-28) [Computer software]. https://CRAN.R-project.org/package=boot

Chambers, C., Akram, S., Adam, V., Pelofi, C., Sahani, M., Shamma, S., & Pressnitzer, D. (2017). Prior context in audition informs binding and shapes simple features. *Nature Communications, 8*(1), 15027. https://doi.org/10.1038/ncomms15027

Crozier, J. B. (1997). Absolute pitch: Practice makes perfect, the earlier the better. *Psychology of Music, 25*(2), 110–119. https://doi.org/10.1177/0305735697252002

Cuddy, L. L. (1968). Practice effects in the absolute judgment of pitch. *The Journal of the Acoustical Society of America, 43*(5), 1069–1076. https://doi.org/10.1121/1.1910941

Darwin, C. J., Turvey, M. T., & Crowder, R. G. (1972). An auditory analogue of the sperling partial report procedure: Evidence for brief auditory storage. *Cognitive Psychology, 3*(2), 255–267. https://doi.org/10.1016/0010-0285(72)90007-2

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods, 47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

Deutsch, D. (1972a). Effect of repetition of standard and of comparison tones on recognition memory for pitch. *Journal of Experimental Psychology, 93*(1), 156–162. https://doi.org/10.1037/h0032496

Deutsch, D. (1972b). Mapping of Interactions in the pitch memory store. *Science, 175*(4025), 1020–1022. https://doi.org/10.1126/science.175.4025.1020

Deutsch, D. (1982). The influence of melodic context on pitch recognition judgment. *Perception & Psychophysics, 31*(5), 407–410. https://doi.org/10.3758/BF03204849

Deutsch, D. (2013). Absolute Pitch. In *The Psychology of Music* (pp. 141–182). Elsevier. https://doi.org/10.1016/B978-0-12-381460-9.00005-5

Deutsch, D., & Roll, P. L. (1974). Error patterns in delayed pitch comparison as a function of relational context. *Journal of Experimental Psychology, 103*(5), 1027–1034. https://doi.org/10.1037/h0037359

Deutsch, D., Henthorn, T., Marvin, E., & Xu, H. (2006). Absolute pitch among American and Chinese conservatory students: Prevalence differences, and evidence for a speech-related critical period. *The Journal of the Acoustical Society of America, 119*(2), 719. https://doi.org/10.1121/1.2151799

Deutsch, D., Dooley, K., Henthorn, T., & Head, B. (2009). Absolute pitch among students in an American music conservatory: Association with tone language fluency. *The Journal of the Acoustical Society of America, 125*(4), 2398–2403. https://doi.org/10.1121/1.3081389

Dewar, K. M., Cuddy, L. L., & Mewhort, D. J. (1977). Recognition memory for single tones with and without context. *Journal of Experimental Psychology: Human Learning and Memory, 3*(1), 60–67. https://doi.org/10.1037/0278-7393.3.1.60

Gervain, J., Vines, B. W., Chen, L. M., Seo, R. J., Hensch, T. K., Werker, J. F., & Young, A. H. (2013). Valproate reopens critical-period learning of absolute pitch. *Frontiers in Systems Neuroscience, 7*, 102. https://doi.org/10.3389/fnsys.2013.00102

Gregersen, P. K., Kowalsky, E., Kohn, N., & Marvin, E. W. (1999). Absolute pitch: Prevalence, ethnic variation, and estimation of the genetic component. *The American Journal of Human Genetics, 65*(3), 911–913. https://doi.org/10.1086/302541

Hartman, E. B. (1954). The influence of practice and pitch-distance between tones on the absolute identification of pitch. *The American Journal of Psychology, 67*(1), 1. https://doi.org/10.2307/1418067

Kim, S., & Knösche, T. R. (2016). Intracortical myelination in musicians with absolute pitch: Quantitative morphometry using 7-T MRI. *Human Brain Mapping, 37*(10), 3486–3501. https://doi.org/10.1002/hbm.23254

Kim, S., & Knösche, T. R. (2017). On the perceptual subprocess of absolute pitch. *Frontiers in Neuroscience, 11*, 557. https://doi.org/10.3389/fnins.2017.00557

Kraft, J. M., & Brainard, D. H. (1999). Mechanisms of color constancy under nearly natural viewing. *Proceedings of the National Academy of Sciences, 96*(1), 307–312. https://doi.org/10.1073/pnas.96.1.307

Krumhansl, C. L. (1979). The psychological representation of musical pitch in a tonal context. *Cognitive Psychology, 11*(3), 346–374. https://doi.org/10.1016/0010-0285(79)90016-1

Krumhansl, C. L., & Castellano, M. A. (1983). Dynamic processes in music perception. *Memory & Cognition, 11*(4), 325–334. https://doi.org/10.3758/BF03202445

Lenth, R. V. (2021). *emmeans: Estimated marginal means, aka least-squares means* (R package version 1.7.0) [Computer software]. https://CRAN.R-project.org/package=emmeans

Levitin, D. J., & Rogers, S. E. (2005). Absolute pitch: Perception, coding, and controversies. *Trends in Cognitive Sciences, 9*(1), 26–33. https://doi.org/10.1016/j.tics.2004.11.007

Levitin, D. J., & Zatorre, R. J. (2003). On the nature of early music training and absolute pitch: A Reply to Brown, Sachs, Cammuso, and Folstein. *Music Perception, 21*(1), 105–110. https://doi.org/10.1525/mp.2003.21.1.105

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49*(2), 433–442. https://doi.org/10.3758/s13428-016-0727-z

Lockhead, G. R., & Byrd, R. (1981). Practically perfect pitch. *The Journal of the Acoustical Society of America, 70*(2), 387–389. https://doi.org/10.1121/1.386773

Lotto, R. B., & Purves, D. (2000). An Empirical Explanation of Color Contrast. *Proceedings of the National Academy of Sciences, 97*(23), 12834–12839. https://doi.org/10.1073/pnas.210369597

Lundin, R. W. (1963). Can perfect pitch be learned? *Music Educators Journal, 49*(5), 49–51. https://doi.org/10.2307/3389949

Macmillan, N. A., & Creelman, C. D. (2004). *Detection Theory* ((0 ed.). ed.). Psychology Press. https://doi.org/10.4324/9781410611147

Makowski, D. (2018). The psycho package: An efficient and publishing-oriented workflow for psychological science. *The Journal of Open Source Software, 3*(22), 470. https://doi.org/10.21105/joss.00470

Marmel, F., Tillmann, B., & Dowling, W. J. (2008). Tonal expectations influence pitch perception. *Perception & Psychophysics, 70*(5), 841–852. https://doi.org/10.3758/PP.70.5.841

Miyazaki, K. (1988). Musical pitch identification by absolute pitch possessors. *Perception & Psychophysics, 44*(6), 501–512. https://doi.org/10.3758/BF03207484

Miyazaki, K. (1989). Absolute pitch identification: Effects of timbre and pitch region. *Music Perception, 7*(1), 1–14. https://doi.org/10.2307/40285445

Miyazaki, K., Makomaska, S., & Rakowski, A. (2012). Prevalence of absolute pitch: A comparison between Japanese and Polish music students. *The Journal of the Acoustical Society of America, 132*(5), 3484–3493. https://doi.org/10.1121/1.4756956

Nguyen, B. N., & McKendrick, A. M. (2016). Visual contextual effects of orientation, contrast, flicker, and luminance: All are affected by normal aging. *Frontiers in Aging Neuroscience, 8*, 79. https://doi.org/10.3389/fnagi.2016.00079

Oxenham, A. J. (2012). Pitch perception. *Journal of Neuroscience, 32*(39), 13335–13338. https://doi.org/10.1523/JNEUROSCI.3815-12.2012

Reymore, L., & Hansen, N. C. (2020). A theory of instrument-specific absolute pitch. *Frontiers in Psychology, 11*, 560877. https://doi.org/10.3389/fpsyg.2020.560877

Russo, F. A., Windell, D. L., & Cuddy, L. L. (2003). Learning the "special note": Evidence for a critical period for absolute pitch acquisition. *Music Perception, 21*(1), 119–127. https://doi.org/10.1525/mp.2003.21.1.119

Schellenberg, E. G., & Trehub, S. E. (2003). Good pitch memory is widespread. *Psychological Science, 14*(3), 262–266. https://doi.org/10.1111/1467-9280.03432

Sergeant, D. (1969). Experimental Investigation of absolute pitch. *Journal of Research in Music Education, 17*(1), 135–143. https://doi.org/10.2307/3344200

Smith, G. (2005). On construct validity: Issues of method and measurement. *Psychological Assessment, 17*(4), 396–408. https://doi.org/10.1037/1040-3590.17.4.396

Takeuchi, A. H., & Hulse, S. H. (1993). Absolute pitch. *Psychological Bulletin, 113*(2), 345–361. https://doi.org/10.1037/0033-2909.113.2.345

Temperley, D., & de Clercq, T. (2013). Statistical analysis of harmony and melody in rock music. *Journal of New Music Research, 42*(3), 187–204. https://doi.org/10.1080/09298215.2013.788039

Van Hedger, S. C., & Nusbaum, H. C. (2018). Individual differences in absolute pitch performance: Contributions of working memory, musical expertise, and tonal language background. *Acta Psychologica, 191*, 251–260. https://doi.org/10.1016/j.actpsy.2018.10.007

Van Hedger, S. C., Heald, S. L. M., Koch, R., & Nusbaum, H. C. (2015a). Auditory working memory predicts individual differences in absolute pitch learning. *Cognition, 140*, 95–110. https://doi.org/10.1016/j.cognition.2015.03.012

Van Hedger, S. C., Heald, S. L. M., & Nusbaum, H. C. (2015b). The effects of acoustic variability on absolute pitch categorization: Evidence of contextual tuning. *The Journal of the Acoustical Society of America, 138*(1), 436–446. https://doi.org/10.1121/1.4922952

Van Hedger, S. C., Heald, S. L., & Nusbaum, H. C. (2018a). Long-term pitch memory for music recordings is related to auditory working memory precision. *Quarterly Journal of Experimental Psychology, 71*(4), 879–891. https://doi.org/10.1080/17470218.2017.1307427

Van Hedger, S. C., Heald, S. L. M., Uddin, S., & Nusbaum, H. C. (2018b). A note by any other name: Intonation context rapidly changes absolute note judgments. *Journal of Experimental Psychology: Human Perception and Performance, 44*(8), 1268–1282. https://doi.org/10.1037/xhp0000536

Van Hedger, S. C., Heald, S. L. M., & Nusbaum, H. C. (2019). Absolute pitch can be learned by some adults. *PLoS One, 14*(9), e0223047. https://doi.org/10.1371/journal.pone.0223047

Van Hedger, S. C., Veillette, J., Heald, S. L. M., & Nusbaum, H. C. (2020). Revisiting discrete versus continuous models of human behavior: The case of absolute pitch. *PLoS One, 15*(12), e0244308. https://doi.org/10.1371/journal.pone.0244308

Van Hedger, S. C., Johnsrude, I. S., & Batterink, L. J. (2022). Musical instrument familiarity affects statistical learning of tone sequences. *Cognition, 218*, 104949. https://doi.org/10.1016/j.cognition.2021.104949

Vanzella, P., & Schellenberg, E. G. (2010). Absolute pitch: Effects of timbre on note-naming ability. *PLoS One, 5*(11), e15449. https://doi.org/10.1371/journal.pone.0015449

Vuvan, D. T., Podolak, O. M., & Schmuckler, M. A. (2014). Memory for musical tones: The impact of tonality and the creation of false memories. *Frontiers in Psychology, 5*, 582. https://doi.org/10.3389/fpsyg.2014.00582

Warren, J. D., Uppenkamp, S., Patterson, R. D., & Griffiths, T. D. (2003). Separating pitch chroma and pitch height in the human brain. *Proceedings of the National Academy of Sciences, 100*(17), 10038–10042. https://doi.org/10.1073/pnas.1730682100

Wong, Y. K., Lui, K. F. H., Yip, K. H. M., & Wong, A. C.-N. (2020a). Is it impossible to acquire absolute pitch in adulthood? *Attention, Perception, & Psychophysics, 82*(3), 1407–1430. https://doi.org/10.3758/s13414-019-01869-3

Wong, Y. K., Ngan, V. S., Cheung, L. Y., & Wong, A. C.-N. (2020b). Absolute pitch learning in adults speaking non-tonal languages. *Quarterly Journal of Experimental Psychology, 73*(11), 1908–1920. https://doi.org/10.1177/1747021820935776

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics, 79*(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2

Wylie, G. R., Yao, B., Sandry, J., & DeLuca, J. (2021). Using signal detection theory to better understand cognitive fatigue. *Frontiers in Psychology, 11*, 579188. https://doi.org/10.3389/fpsyg.2020.579188