Memory & Cognition
https://doi.org/10.3758/s13421-023-01429-z

Absolute pitch judgments of familiar melodies generalize across timbre and octave

Stephen C. Van Hedger[1,2] · Noah R. Bongiovanni[3] · Shannon L. M. Heald[4,5] · Howard C. Nusbaum[4,5]

Accepted: 24 April 2023

Abstract

Most listeners can determine when a familiar recording of music has been shifted in musical key by as little as one semitone (e.g., from B to C major). These findings appear to suggest that absolute pitch memory is widespread in the general population. However, the use of familiar recordings makes it unclear whether these findings genuinely reflect absolute melody-key associations for at least two reasons. First, listeners may be able to use spectral cues from the familiar instrumentation of the recordings to determine when a familiar recording has been shifted in pitch. Second, listeners may be able to rely solely on pitch *height* cues (e.g., relying on a feeling that an incorrect recording sounds "too high" or "too low"). Neither of these strategies would require an understanding of pitch chroma or musical key. The present experiments thus assessed whether listeners could make accurate absolute melody-key judgments when listening to *novel* versions of these melodies, differing from the iconic recording in timbre (Experiment 1) or timbre and octave (Experiment 2). Listeners in both experiments were able to select the correct-key version of the familiar melody at rates that were well above chance. These results fit within a growing body of research supporting the idea that most listeners, regardless of formal musical training, have robust representations of absolute pitch – based on pitch chroma – that generalize to novel listening situations. Implications for theories of auditory pitch memory are discussed.

Keywords Absolute pitch · Music cognition · Memory · Generalization

Introduction

Music is omnipresent in everyday life. Regardless of one's affinity toward music, it is difficult to navigate daily activities (e.g., travel, work, shopping) without hearing recordings of popular songs. Adult listeners spend several hours per week listening to music (Greasley & Lamont, 2011), and music has been estimated to be present in over one-third of our waking hours (Juslin et al., 2008). Given that digitized recordings are almost always heard at the same pitch level, this means that listeners have the opportunity to implicitly learn associations between specific recordings and musical keys. Although a growing body of research has indeed suggested that most listeners can tell when a familiar recording has been altered in pitch (e.g., Jakubowski et al., 2017; Schellenberg & Trehub, 2003; Van Hedger et al., 2018), by using the iconic recording it is unclear whether this kind of *latent pitch memory* is grounded in genuine melody-key associations, as opposed to other pitch cues (e.g., timbral or pitch height cues). By using acoustically novel versions of familiar melodies, the present study directly examines whether latent pitch memory for familiar melodies is based on an implicit understanding of its associated musical key.

Addressing the question of whether listeners have genuine melody-key associations has notable implications for linking this widespread latent pitch memory with the phenomenon of absolute pitch (AP). AP is a rare ability characterized by the ability to produce or identify a musical note without the

✉ Stephen C. Van Hedger
svanhedg@uwo.ca

1 Department of Psychology, Huron University College at Western, 1349 Western Road, London, ON N6G 1H3, Canada

2 Department of Psychology and Brain and Mind Institute, Western University, London, Ontario, Canada

3 Department of Music, University of Notre Dame, South Bend, IN, USA

4 Department of Psychology, University of Chicago, Chicago, IL, USA

5 Center for Practical Wisdom, University Chicago, Chicago, IL, USA

Published online: 10 May 2023

use of an external reference (e.g., Takeuchi & Hulse, 1993, Bachem, 1955). AP possessors, unlike the general population, form explicit categories based on pitch *chroma* – the quality of a specific note, regardless of the tonal context, instrumental timbre, or octave in which the pitch is presented (Takeuchi & Hulse, 1993). Non-AP possessors, on the other hand, are thought to predominantly listen to music through a relative pitch (RP) lens, meaning that their perception of a given pitch is influenced largely by the surrounding context of other pitches (Levitin & Rogers, 2005). For example, the ability to differentiate notes based on pitch *height* (e.g., understanding a given pitch as "lower" or "higher" than an initial pitch) is present in most adults and even in infants as young as 6 months (Plantinga & Trainor, 2005). Although listeners might develop long-term memory representations based on pitch height (e.g., knowing that a flute plays higher than a tuba without necessarily having to hear a reference sound from a tuba), this form of pitch memory is considered to be distinct from "genuine" AP representations based on pitch chroma (e.g., Bachem, 1937; Kim & Knösche, 2016, 2017), in part because it involves coarse judgments based on extreme differences in pitch height. However, studies in the literature clarify that AP possessors do rely at least in small part on other mechanisms when making their pitch judgments. For example, among AP possessors, performance has been observed to vary as a function of several factors, including instrumental familiarity, harmonic complexity, octave register, and even whether the notes are sung versus produced non-vocally (see Lockhead & Byrd, 1981; Miyazaki, 1989; Sergeant, 1969; Oxenham, 2012; Vanzella & Schellenberg, 2010). The meaningful variability in performance suggests that AP ability itself exists as a continuum (Levitin & Rogers, 2005); while some AP possessors have very strong constructs for pitch chroma, others may be somewhat weaker, such that one would see a greater reliance on other acoustical cues. It is important to emphasize that even though evidence suggests AP possessors use all auditory information available to them in musical contexts to identify pitches, the judgments are based fundamentally in representations of *chroma*, and the degree of reliance on these secondary cues is far less than absolute, and varies situationally.

Despite the fact that the vast majority of people do not have AP, an accumulating body of research has shown that most listeners have remarkably fine-grained memory representations for absolute pitches, at least when tested in more ecologically valid contexts. Participants given a production task asking them to sing the first notes of familiar melodies showed high levels of consistency in starting pitch, even when experimental sessions took place days apart (Halpern, 1989). Levitin (1994) found that when individuals were asked to hum a few bars from a self-selected well-known recording, they started on the correct absolute pitch at a rate significantly higher than chance. A more recent large-scale follow-up study replicated the findings of Levitin (1994), although reported overall lower effect sizes than the original study (Frieler et al., 2013). Taken together, these findings suggest that individuals are consistent in their production of musical melodies and spontaneously reproduce familiar melodies in the correct musical key at a rate that is significantly above chance.

In addition to these production-based tasks involving singing or humming, several studies show that listeners also demonstrate robust absolute pitch memory for familiar recordings when tested in perceptually based paradigms. For example, in an early investigation of absolute pitch memory for familiar pieces of music, Terhardt and Seewann (1983) found that, when presented with two versions of excerpts from Bach's *The Well-Tempered Clavier*, listeners were able to select the nominally correct version (i.e., the version in the correct musical key) versus alternatives that were shifted by as little as one semitone. In the same study, the researchers modified the piano recordings so that the pieces were presented again with a computerized, synthetic timbre. They found similar performance despite the timbral manipulation, indicating that the level of accuracy achieved by the musicians was due to pitch and not another cue, such as familiarity with the particular instrument. However, the participants in Terhardt and Seewann (1983) were all highly trained musicians, with 11 possessing AP. As such, from these findings it is unclear whether this generalization across instrumental timbre is a general feature of absolute pitch memory for familiar recordings, or whether this kind of generalization is limited to individuals with significant musical training. Subsequent research has clearly shown that absolute pitch memory for well-known recordings does not require extensive musical training (e.g., Jakubowski et al., 2017; Schellenberg & Trehub, 2003; Van Hedger et al., 2018). However, these studies have tested pitch memory by presenting participants with iconic recordings of familiar melodies (e.g., the same recording that one would hear outside of the experiment, when listening to the radio or watching television). As a result, these studies do not clarify how broadly listeners can judge and remember familiar melodies based on pitch chroma specifically, as opposed to other sources of information such as instrumentation or pitch height. In order to more conclusively determine whether listeners are using pitch chroma in these judgments, it is necessary to systematically eliminate these additional sources of information.

To investigate the extent to which pitch memory for familiar melodies reflects an understanding of pitch chroma, the present study assesses how pitch memory for familiar melodies (e.g., taken from popular songs heard on the radio) generalizes across *both* timbre and octave. Although prior research has demonstrated timbre generalization using a similar paradigm (Terhardt & Seewann, 1983), it is unclear whether these findings generalize beyond individuals with significant musical training. Moreover, even if timbre generalization is widespread in the population, this does not mean that participants are necessarily relying on pitch chroma, as they still could be using pitch height to make judgments. Similar to theoretic approaches to operationalizing genuine AP, we reason that the inclusion of both timbre and octave generalization is a critical test of pitch chroma (e.g., see Bongiovanni et al., 2023).

Across two experiments, we adopt a forced-choice paradigm, in which participants hear two versions of a familiar melody and judge which one sounds correct (i.e., more similar to the version heard outside of the experiment). In Experiment 1, we specifically investigated how timbral cues influence pitch memory by asking participants to judge novel, instrumental versions of popular melodies. In Experiment 2, we investigated how both timbral and pitch height cues influence pitch memory by presenting novel versions of popular melodies, presented either in the correct octave register or shifted in register by one octave (either up or down). If participants' memories for popular songs are grounded in an implicit understanding of pitch chroma (i.e., an implicit understanding of the musical key of the song), then they should be able to generalize across timbre and octave and select the correct version above chance. In contrast, if participants in previous studies were relying on auditory cues outside of chroma (e.g., instrumentation or a general sense of pitch height), then performance is expected to be at chance in both experiments. Overall, these two experiments aid our understanding of the nature of pitch memory representation among the general population, and specifically among non-AP possessors without musical training.

Experiment 1

Method

Participants

Fifty-eight University of Chicago undergraduate students were recruited for the experiment. Three participants self-reported possessing AP and were thus excluded from the primary analyses, leaving 55 participants (M_{age} = 19.49 years, *SD* = 1.15, range 18–24 years). There were 28 participants in the "Perceptually Rich" condition and 27 participants in the "Perceptually Sparse" condition. Sample size was based on prior work examining pitch memory for popular recordings among adults – specifically, Experiment 1 of Schellenberg and Trehub (2003). This experiment, which assessed pitch memory for highly familiar songs similar to the present experiment, found a large effect size (d = 1.16). Given that such a large effect size was considered unlikely given the acoustic manipulations of the present experiment, we ensured that each condition was adequately powered (> .80) to detect deviations from chance with an anticipated moderate effect size (d = 0.58), half of that found by Schellenberg and Trehub (2003). Participants were not specifically recruited for their musical backgrounds; however, a majority (89.1%) of participants reported at least some formal musical training (*M* = 5.87 years, *SD* = 4.38, range 0–18 years). All participants provided informed consent and were compensated with course credit. The research protocol was approved by the University of Chicago Institutional Review Board.

Materials

The 40 popular recordings were the same as those reported by Van Hedger et al. (2018) and consisted of 20 popular songs (e.g., Billboard Top 40) and 20 popular movie and television themes (see Table S1 in Online Supplemental Material (OSM)). A pilot group of ten University of Chicago undergraduates (who did not participate in the Experiment 1) rated these 40 recordings as highly familiar, providing a mean rating of 4.11 (*SD* = 0.53) on a Likert scale ranging from 1 (*not at all familiar)* to 5 (*extremely familiar*).

The stimuli for the "Perceptually Rich" condition were selected by searching for skillful cover versions of the popular recordings on YouTube. The key of each recording was verified to match with the iconic recording by the authors. For example, the cover version of "Piano Man" by Billy Joel was verified to be in the key of C major, as this is the key from the iconic recording of the song. Similar to Van Hedger et al. (2018), each cover was of variable length (*M* = 11.18 s, *SD* = 2.83 s). The variable length was to ensure that the recording did not cut off a musical phrase. The pitch of each cover was manipulated in Audacity (Audacity Team, 2021) using the built-in "Change Pitch" function, resulting in five versions of each cover recording (correct pitch, ± 1 semitone shift, and ± 2 semitone shift). To prevent participants from using audio quality as a cue for selecting the correct version of each recording, *all* recordings were run through the pitch-shifting algorithm twice. For example, stimuli that were shifted in pitch by two semitones were shifted by ± 1 semitone two times, stimuli that were shifted in pitch by one semitone were shifted by ± 0.5 semitone two times, and stimuli that were presented in the correct key were either shifted by +1 followed by -1 semitones (or -1 followed by +1 semitones). Although this pitch-shifting approach was meant to minimize auditory quality differences across the different versions of the songs, it is possible that there were still some small perceptible differences that could have facilitated participants choosing the correct version of each song. However, prior research using this pitch-shifting algorithm (e.g., Schellenberg & Trehub, 2003; Van Hedger et al., 2018) has shown that participants are statistically at chance when judging unfamiliar songs, which suggests that any auditory quality differences that may exist across song versions are minimal and do not influence the accuracy of judgments. Each rich cover stimulus was digitized at 44.1 kHz with 16-bit depth and normalized to -20 dB Full Spectrum.

The stimuli for the "Perceptually Sparse" condition were created by the first author and a research assistant using Reason music production software (Propellerhead, Stockholm, Sweden). Each stimulus was a monophonic (i.e., one-note-at-a-time) representation of the melodic line from each recording, recorded with a grand piano timbre. The sparse cover stimuli were also of variable length and were longer than the rich cover stimuli ($M = 18.92$ s, $SD = 4.98$ s), which was done to allow participants a greater amount of time to recognize the melody given the sparse nature of the stimuli. Given that the recordings were represented in MIDI format, the sparse covers were simply transposed in key by one or two semitones prior to exporting as an audio file. Five versions of each sparse cover were created (correct key, ± 1 semitone shift, and ± 2 semitone shift). Similar to the rich covers, each sparse cover stimulus was digitized at 44.1 kHz with 16-bit depth and normalized to -20 dB Full Spectrum.

The experiment was programmed in E-Prime 2.0 (Psychology Software Tools, Sharpsburg, PA, USA). Participants listened to the sounds through Sennheiser HD570 circumaural headphones.

Procedure

After providing informed consent, participants were introduced to the main task. Participants were specifically instructed that they would be hearing "cover" versions of popular musical melodies. Specifically, participants were told that they would hear two cover versions of a melody and would have to judge whether the first or the second was correct. Correctness was defined as sounding more like the "original" (iconic) version of the melody. To ensure that participants were familiar with what was meant by "cover" versions, participants were presented with an example stimulus from a song not tested in the main task.

Participants then completed the main task, which consisted of 40 trials. On each trial, participants heard the two versions of the cover recording and then made a forced-choice judgment as to whether the first or the second version was correct. Participants were additionally asked to rate their familiarity with the melody on a 5-point scale, ranging from 1 (*not at all familiar*) to 5 (*extremely familiar*). The incorrect version of the melody could either be one semitone too high (25% of trials), two semitones too high (25% of trials), one semitone too low (25% of trials), or two semitones too low (25% of trials). These incorrect trial types were not randomly assigned to melodies, but rather were pseudo-randomly determined through different counterbalanced versions of the experiment. Specifically, there were four versions of the experiment, counterbalanced across participants, to ensure that each melody was heard with each incorrect version across all participants. Trial order in the main task was randomized, and ordering of the correct stimulus (first vs. second) was counterbalanced.

Following the main task, participants completed a basic demographic and musical experience questionnaire. The demographic questionnaire assessed participants' age and gender. From the music experience questionnaire, we extracted (1) whether participants still actively played a musical instrument, (2) the reported number of years of musical training, and (3) the extent to which participants listened to popular music, expressed as a percentage of their total music listening. After the questionnaire, participants were provided with a debriefing form and given course credit.

Data analysis

Before performing the primary analyses, we first removed trials in which participants reported no familiarity with the melody. This decision was justified because this form of pitch memory relies on at least some prior familiarity with the iconic recording (e.g., see Schellenberg & Trehub, 2003). On average, participants reported at least some prior familiarity on a majority of trials (73.7%), which is relatively high considering (1) the novel nature of the recordings used in the experiment and (2) the fact that participants were not provided with any explicit information about the melody (e.g., song title or artist) during the experiment.

To assess whether participants were above chance in identifying the "correct" version of the cover recording, we used one-sample t-tests and Bayesian equivalents, calculated in JASP (Marsman & Wagenmakers, 2017). These analyses tested mean accuracy against a 50% chance estimate. The Bayesian analysis provided a Bayes Factor (BF_{10}), which quantifies the relative evidence in favor of the alternative hypothesis compared to the null hypothesis. For example, a BF_{10} of 10 would mean that the alternative hypothesis is 10 times more likely than the null given the data, whereas a BF_{10} of 0.10 would mean that the alternative hypothesis is one-tenth as likely as the null given the data.

We then modeled performance accuracy as a function of (1) condition, (2) self-reported familiarity with the melody, (3) whether the incorrect version was smaller (± 1 semitone) or larger (± 2 semitones) in size, (4) the number of years of reported musical training, (5) whether a participant actively played a musical instrument, and (6) the extent to which participants listened to popular music. We modeled trial-by-trial data and thus used a generalized linear mixed-effects model with a binomial link (as the dependent response was either 0 or 1 for incorrect or correct) using the "lme4" package (Bates et al., 2015). Participants were modeled with random intercepts. We planned to model the random intercept of melody as well, but singularity warnings suggested that the model was overfit with this term. The significance of the terms in the mixed-effects model was assessed in two primary ways. First, we used the associated *p*-values with each term in the model. Second, we calculated Bayes Factors for each fixed effect in the model. Bayes Factors were calculated by comparing the null model with a model that included the fixed effect.

Results

Testing performance against chance

Overall, participants selected the correct version of the recording 59.7% ($SD = 11.3\%$) of the time, which was significantly higher than the chance estimate of 50%, $t(54) = 6.39$, $p < .001$, d = 0.86, $BF_{10} = 3.27e5$. Additionally, both the rich and sparse cover conditions were independently above chance. Mean accuracy in the rich cover condition was 61.9% ($SD = 11.0\%$), $t(27) = 5.75$, $p < .001$, d = 1.09, $BF_{10} = 4848$. Mean accuracy in the sparse condition was 57.4% ($SD = 11.3\%$), $t(26) = 3.40$, $p = .002$, d = 0.65, $BF_{10} = 17.10$. Figure 1 represents overall performance across both rich and sparse cover conditions.

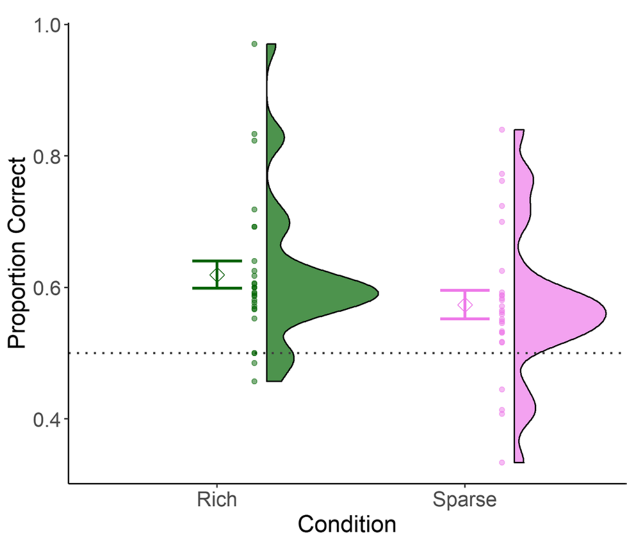

Fig. 1 Accuracy for both rich and sparse cover conditions in Experiment 1. *Note:* Each condition displays the mean and standard error (left), individual participant points (middle), and density plots of how performance was distributed (right). The dotted horizontal line represents chance performance

Predicting performance based on measured factors

Table 1 provides a summary of each fixed-effect term from the mixed-effects analyses. In the full model, there were no significant effects. The effect of condition (sparse vs. rich covers) was marginally significant; however, the Bayes Factor analysis did not support the inclusion of this term compared to an intercept-only null model.

Discussion

Experiment 1 provides evidence that listeners can accurately judge the musical key of novel instances of familiar melodies. In both the rich and sparse cover conditions, participants were well above chance in selecting the "correct" version of the familiar melody, which is notable considering the specific recordings were very likely novel to participants in the rich cover condition and guaranteed to be novel in the sparse condition (given that they were created specifically for the present experiment). Moreover, it is notable that participants' performance in both conditions was statistically comparable, given that the sparse covers were monophonic renditions of the familiar recordings and thus was essentially stripped of any additional harmonic or timbral cues that could have aided listeners. Thus, the above-chance performance observed for the sparse covers in particular suggest that listeners have a more generalized association of melodies with specific musical keys and can make absolute melody-key (chroma) judgments in novel and controlled listening conditions. Taken together, Experiment 1 demonstrates that absolute memory for popular melodies cannot be entirely explained via an alternative, non-absolute pitch account such as sensitivity to shifted spectral information.

Table 1 Results of the mixed-effects models from Experiment 1

Term	*B*	*SE*	*p*	BF_{01}
Condition	-0.21	0.12	.089	5.74
Familiarity	0.03	0.05	.534	14.02
Shift Magnitude	-0.14	0.10	.184	7.66
Music Training	0.04	0.07	.521	7.05
Active Musician	0.26	0.17	.135	3.92
Pop Music Listening	0.14	0.24	.570	9.26

The beta coefficients (B), standard errors (SE), and p-values (p) are all taken from a full model that included all fixed effects. In contrast, the Bayes Factors (BF_{01}) were calculated by comparing a null model with a model that contains each fixed effect considered separately. Fr example, the BF01 for Condition was calculated by comparing an intercept-only model to a model containing both intercept and Condition. For interpretability, we report BF_{01}, which provides the relative evidence in favor of the null model compared to the model with the fixed-effect term. Random intercepts of participants were modeled in all cases

Yet, Experiment 1 cannot conclusively rule out the possibility that participants were using non-chroma-related strategies in making their judgments. Most notably, the cover recordings in both the rich and sparse conditions were almost exclusively in the same octave as the original, iconic recording. Consequently, the fundamental frequencies of the melodic lines in the present experiment were largely identical to the iconic recordings. This is problematic for interpreting the results in terms of absolute pitch representations because pitch height and chroma cannot be disentangled. Critically, if these absolute pitch representations are similar to those observed in genuine AP, then listeners should be able to make accurate judgments of familiar melodies even when they are shifted in octave from the original, iconic recordings (as musical key is based on pitch chroma and independent of pitch height). Alternatively, if participants cannot make absolute judgments when the melodies are shifted in octave, this suggests a different kind of pitch representation based on pitch height and not chroma. Experiment 2 tests these possibilities by systematically varying the octave in which familiar melodies are presented across participants.

Experiment 2

Method

Participants

We recruited 200 participants from Amazon Mechanical Turk and 183 were included in the primary analyses (M_{age} = 34.04 years, SD = 9.81, range 20–65 years). Participants were excluded if they failed the auditory attention check (n = 9), self-reported possessing AP (n = 4), or did not recognize at least 50% of the melodies (n = 5). In total, 60 participants were included in the standard-octave condition, 63 participants were in the high-octave condition, and 60 participants were in the low-octave condition. Sample size was initially determined based on an a priori power analysis outlined in the preregistration, in which we determined that a sample size of 50 participants per condition would result in statistical power of .90 assuming a mean accuracy of 55% (which was considered to be the smallest meaningful effect of interest, as it represents exactly one trial above chance, 11 of 20). We slightly increased the sample size relative to the preregistration plan (see section Deviations from preregistration) in part to ensure that, after participant exclusions, we would have a minimum of 50 participants in each condition. Participants had variable amounts of musical training, with 57.9% reporting at least some musical training and only 18.0% reporting actively playing music. Although the mean number of years of reported musical training was 4.26 years (SD = 7.71 years), the median and modal response was zero years of musical training, suggesting that a small number of participants reported high amounts of musical training. All participants had a minimum 90% prior approval rating and a minimum of 50 prior completed assignments on Mechanical Turk. The protocol was approved by the University of Chicago Institutional Review Board.

Materials

We selected 20 melodies for use in the present experiment. The melodies were a subset of the 40 melodies used in Experiment 1 (see Table S2 in the OSM for a complete list). The total number of tested melodies was reduced in an effort to keep the experiment relatively short given the online sample. Similar to the sparse cover condition from Experiment 1, each melody was recorded monophonically using a grand piano timbre. Each melody was represented in MIDI format, meaning that the key shifts and octave shifts could be done via transposing the MIDI file prior to exporting it as an audio file. This is especially advantageous given the present approach of varying octave, in which such extreme shifts in auditory frequency would be more likely to result in noticeable auditory artifacts. Given the approach of the present experiment, there were nine versions of each melody, as melodies could be presented in three octaves (standard, high, low) and, within each octave condition, each melody had three versions (correct key, +1 semitone, -1 semitone). Unlike Experiment 1, all melodies had a fixed duration of 10 s with a 500-ms linear fade out. All recordings were digitized at 44.1 kHz with 16-bit depth and normalized to -20 dB Full Spectrum. The experiment was programmed in jsPsych 6 (de Leeuw, 2015).

Procedure

Upon clicking on the experiment link, participants were randomly assigned to either the standard, higher, or lower octave condition. Participants first provided informed consent by pressing a keyboard button acknowledging that they had read the consent form and agreed to participate in the experiment. Following the consent procedure, participants completed a short auditory calibration. This calibration involved playing a 30-s noise and allowing participants to adjust their computer's volume to ensure the noise, which was normalized to the same level as the melodies, was being played at a comfortable volume.

Following the auditory calibration, participants were introduced to the main task. Participants were instructed that they would hear two versions of well-known melodies, played on a piano, and would have to determine which version was "correct." Participants were not specifically alerted to the between-participant manipulation of octave. Rather, participants were told that one of the two versions would always be correct and that they should rely on their best guess if they were not sure.

The main pitch judgment task had the same structure regardless of octave condition. There were 20 total trials, representing the 20 melodies used in the experiment (presented in a randomized order). Each trial began with a written prompt of the melody participants would hear (e.g., "You will now hear two versions of [Artist's] [Song Title]"). We decided to provide participants with information about the melody because we did not want participants to spend time trying to determine whether the melody was familiar to them. Additionally, providing a written cue about the melody was not hypothesized to influence absolute pitch memory, as it in no way alerted participants as to which version was correct. After participants received the information about the melody, they pressed a button to initiate the playing of both melodies. Participants heard two versions of each melody, with the correct version appearing in the first position 50% of the time. The incorrect version of the melody was either one semitone too high (50% of trials) or one semitone too low (50% of trials). Following the presentation of both melodies, participants were asked to make a forced-choice judgment about which version sounded correct. Participants then rated their familiarity with the melody by clicking on one of five possible options (*not at all, slightly, somewhat, moderately, extremely*). If a participant responded with *not at all*, the trial was discarded from analysis. If a participant responded with *not at all* to ten or more trials, they were excluded from subsequent analyses.

Immediately following the main pitch judgment task, participants were given an auditory attention check. Participants heard an auditory prompt to click on one of three marked buttons on the screen. There were two auditory attention trials, and participants had to pass both to be included in the main analyses.

Following the auditory attention check, participants completed a short questionnaire. In this questionnaire, participants were asked to provide basic demographic information (age, gender, and ethnicity), as well as information related to their musical background. All participants were asked whether they had ever played a musical instrument (voice included), whether they currently play a musical instrument, and whether they possess absolute pitch. Participants who reported ever playing a musical instrument were asked additional questions about the number of years of formal musical training they received, their primary musical instrument, and the age at which they first began musical instruction. If a participant reported possessing AP, they were excluded from the analyses. Participants were additionally asked about the strategies they used during the main task. Options included "Quietly humming / singing along with the recording," "Imagining humming / singing along with the recording," "Imagining the 'correct' recording playing in your mind's ear," and "Relying on a 'gut' feeling." Participants provided a yes/no response to each of these potential strategies and could select more than one. Additionally, participants were given the space to outline a different strategy they used that was not listed. Following the questionnaire, participants were given a unique completion code, given information about the purpose of the study, and paid for their participation.

Data analysis

Testing performance against chance To test whether participants were above the chance estimate of 50%, we used one-sample t-tests and Bayesian equivalents, calculated in JASP. We also non-parametrically assessed how many participants had a mean accuracy that was above 50% using a non-parametric binomial test and Bayesian equivalent, conceptually similar to prior reports of AP memory for familiar melodies (Schellenberg & Trehub, 2003). For the binomial test, we excluded participants who scored exactly 50% and assessed the relative proportion of remaining participants who were above chance (vs. below chance). We conducted these analyses separately for each octave condition (standard, high, low) to assess whether each condition was independently above chance.

Assessing effects of octave condition (ANOVA) To test whether participants in the octave-shifted conditions (i.e., high, low) differ from participants in the standard-octave condition, we first implemented a one-way analysis of variance (ANOVA) and Bayesian equivalent, with mean accuracy as the dependent variable and octave group as the independent variable.

Additionally, in order to disentangle pitch height and chroma contributions to performance, we constructed a two-way mixed ANOVA and Bayesian equivalent, with mean accuracy as the dependent variable. In this analysis, we only considered shifted octave groups (low vs. high, between-participant factor) and shift direction of the incorrect melody (+1 semitone vs. -1 semitone, within-participant factor). If a significant interaction were observed, post hoc tests assessed if the higher octave leads to better accuracy when the incorrect version is shifted by +1 semitone or, conversely, if the lower octave leads to better accuracy when the incorrect version is shifted by -1 semitone. Such a pattern would indicate that pitch height is influencing participant responses. Further, we determined the degree to which semitone direction accuracy significantly deviates from chance. For example, if participants in the lower-octave condition are significantly above chance when the incorrect melody is flat, but significantly below chance when the incorrect melody is sharp, this would suggest that pitch height is dominant over chroma in making an implicit pitch memory judgment.

Assessing effects of octave condition (GLMM) Finally, we replicated the ANOVA analyses using generalized linear mixed-effects models (GLMMs). In an initial model, we just include octave condition (with the standard octave condition serving as the reference category). Significant effects of octave would reflect significant deviations in pitch memory accuracy compared to the standard octave condition. This model included random intercepts for melody. In a secondary model, we assessed if octave condition interacts with semitone direction by including semitone direction in the model. We only considered participants in the incorrect octave conditions, similar to the above ANOVA. This model included random intercepts for both participant and melody. Finally, for both of these models we added music experience (number of years of playing), melody familiarity, demographic information, and self-reported strategy information (coded as 1 or 0 in terms of whether the strategy was reported) in an exploratory fashion to see if this additional information results in a better-fitting model (assessed via Bayes Factors).

Deviations from preregistration

There were two main deviations from the preregistration of the project. The first was related to the sample size, which was higher ($n = 200$) than what was outlined in preregistration ($n = 150$). The larger sample size was collected because of a change in the availability of funds, to increase statistical power, and updated expectations with regard to the anticipated proportion of participants that would be excluded in an online setting. With a minimum sample size of 60 in each condition, we are adequately powered ($1-\beta = 0.80$) to detect small-to-medium effect sizes (d = 0.325).

The second deviation was related to the program used to code the experiment (jsPsych). This change was implemented due to increased expertise in jsPsych experimental programming by the first author. Ultimately, we decided to use jsPsych because it provides greater flexibility with respect to how it can be served online, and it does not require a specialized license to run over the web.

Results

Testing performance against chance

Mean performance across all conditions is plotted in Fig. 2. Mean performance in the standard-octave condition was 55.1% (SD = 11.7%), which was significantly above the chance estimate of 50%, $t(59) = 3.40$, $p < .001$, d = 0.44, $BF_{10} = 22.61$. This provides a replication of the sparse cover

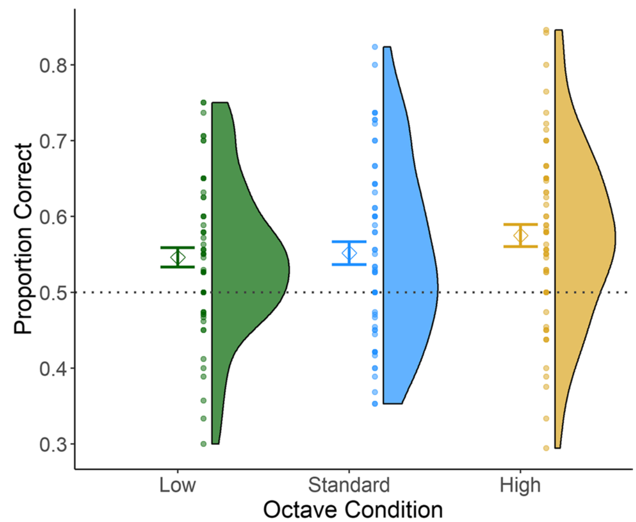

Fig. 2 Accuracy across all three octave conditions in Experiment 2. *Note:* Each condition displays the mean and standard error (left), individual participant points (middle), and density plots of how performance was distributed (right). The dotted horizontal line represents chance performance

condition from Experiment 1. Critically for the hypotheses of the present experiment, we also observed above-chance performance for both of the shifted octave conditions. Mean performance in the higher-octave condition was 57.5% (SD = 11.6%), which was significantly above the chance estimate of 50%, $t(62) = 5.12$, $p < .001$, d = 0.64, $BF_{10} = 5280$. Mean performance in the lower-octave condition was 54.6% (SD = 9.9%), which was significantly above the chance estimate of 50%, $t(59) = 3.57$, $p < .001$, d = 0.46, $BF_{10} = 36.48$.

The non-parametric binomial tests additionally provided evidence that each octave condition was independently above chance. In the standard-octave condition, 34 of 52 analyzable participants (65.4%) scored above chance, $p = .036$, $BF_{10} = 1.99$. In the higher-octave condition, 45 of 58 analyzable participants (77.6%) scored above chance, $p < .001$, $BF_{10} = 1548$. In the lower-octave condition, 37 of 52 analyzable participants (71.2%) scored above chance, $p = .003$, $BF_{10} = 18.96$.

Assessing effects of octave condition (ANOVA)

In the one-way ANOVA, the effect of octave condition was not significant, $F(2,180) = 1.18$, $p = .311$, $\eta^2 = .013$, $BF_{10} = 0.16$. The Bayes Factor suggests that the null model (i.e., the model that does not contain octave condition) is approximately 6.25 times more likely than the model containing octave condition (i.e., 1 / 0.16). Given the nonsignificant effect of octave condition, post hoc analyses were not performed.

The two-way mixed ANOVA, which coded for the incorrect melody shift (-1 semitone, +1 semitone) as a repeated measure and shifted octave condition (low, high) as a between-participant factor, showed a marginally significant main effect of semitone shift, $F(1,121) = 3.08$, $p = .082$, $\eta^2 = .024$, $BF_{Inclusion} = 1.36$. This effect was characterized by overall lower performance when the incorrect version of the melody was -1 semitone ($M = 54.0\%$, $SD = 16.4\%$) as compared to +1 semitone ($M = 58.2\%$, $SD = 17.9\%$). There was also a significant interaction (Fig. 3) between condition and semitone shift, $F(1,121) = 5.81$, $p = .017$, $\eta^2 = .045$, $BF_{Inclusion} = 2.17$, suggesting the relative disruption of hearing a +1 versus -1 semitone foil depended on the octave condition. Consistent with our predictions, participants in the higher-octave condition showed relatively better performance when the incorrect foil was +1 semitone compared to -1 semitone. In contrast, participants in the lower-octave condition displayed the *opposite* pattern, in which performance was better when the foil was -1 semitone compared to +1 semitone.

Follow-up assessments of the condition-by-semitone-shift interaction provided some evidence that participants were not statistically above chance when the foil melody was closer in AP height to the iconic recording (i.e., when the foil was shifted by +1 semitone in the lower-octave condition and -1 semitone in the higher-octave condition). Participants in the lower-octave condition selected the correct melody 55.4% of the time when the foil was shifted by -1 semitone, which was statistically above the chance estimate, $t(59) = 2.64$, $p = .011$, d = 0.34, $BF_{10} = 3.35$. In contrast, participants in the lower-octave condition selected the correct melody 53.9% of the time when the foil was shifted by +1 semitone, which was not statistically above chance, $t(59) = 1.66$, $p = .102$, d = 0.22, $BF_{10} = 0.52$. The effect of semitone shift in the higher-octave condition was even more striking. Participants in the higher-octave condition selected the correct melody 62.2% of the time when the foil was shifted by +1 semitone, which was robustly above chance, $t(62) = 5.75$, $p < .001$, d = 0.72, $BF_{10} = 4.89e4$. In contrast, when the foil was shifted by -1 semitone, participants in the higher-octave condition selected the correct melody 52.7% of the time, which was not above the chance estimate, $t(62) = 1.25$, $p = .216$, d = 0.16, $BF_{10} = 0.29$. In sum, these analyses demonstrate that participants were not statistically above chance when the incorrect "foil" melody was shifted closer in AP height to the iconic recording of the melody.

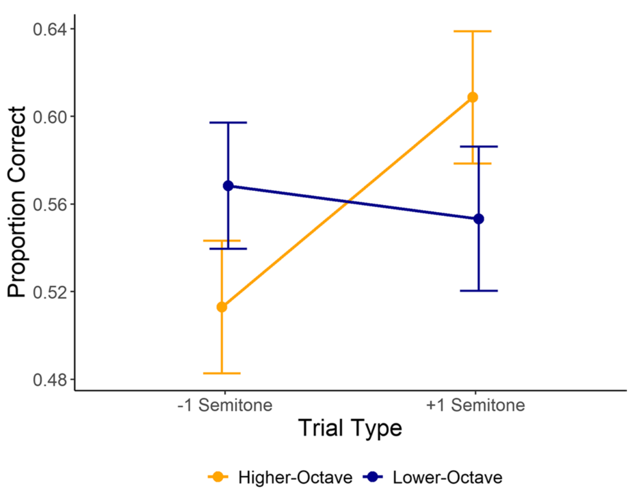

Fig. 3 Interaction of incorrect melody version and octave condition in Experiment 2. *Note:* Error bars represent ± 1 standard error of the mean. Trial type refers to the manner in which the incorrect version of the melody (i.e., the "foil") was shifted – either higher in pitch by one semitone or lower in pitch by one semitone

Assessing effects of octave condition (GLMM)

The GLMMs were conceptually aligned with the ANOVAs. The GLMM that only included octave condition did not show a significant effect of octave condition. Performance in the higher-octave condition, B = 0.12, SE = 0.09, $p = .171$, and performance in the lower-octave condition, B = -0.01, SE = 0.09, $p = .938$, did not differ relative to the standard-octave condition, which was used as the reference level. The expanded model, which included age, years of musical training, self-reported task strategies, and melody familiarity only included one significant term: melody familiarity, B = 0.09, SE = 0.04, $p = .013$. However, the simpler model (i.e., only containing octave condition and the random effects) was still favored over the full model, $BF_{10} = 7.49e\text{-}6$, suggesting that these measures did not explain variance in performance.

The GLMM examining whether performance differed as a function of the incorrect melody shift (-1 semitone vs. +1 semitone) and condition (higher-octave vs. lower-octave) also conceptually replicated the ANOVA. There was a main effect of incorrect melody shift, B = 0.38, SE = 0.12, $p = .002$, with overall higher performance when the melody was shifted by +1 semitone compared to -1 semitone. There was additionally a significant interaction between incorrect melody shift and condition, B = -0.42, SE = 0.17, $p = .015$. The main effect of octave condition was not significant, B = 0.08, SE = 0.12, $p = .498$. Melody familiarity in the expanded model was marginally significant, B = 0.08, SE = 0.04, $p = .061$. All of the other expanded measures (age, years of musical training, self-reported task strategies) were nonsignificant, and the simple model – containing just incorrect melody shift, octave condition, and the random effects – provided a better fit of the data, $BF_{10} = 1.69e\text{-}5$.

Discussion

Experiment 2 demonstrates that participants can generalize across octaves in determining whether a familiar melody is being played in the correct key. Generalization across octaves is an important feature of any memory representation based on pitch chroma (as opposed to pitch height), and therefore the present results demonstrate that pitch memory for well-known melodies is based on an understanding of pitch chroma, similar to the phenomenon of AP (e.g., Takeuchi & Hulse, 1993). Grounding these results in the two-stage model of AP processing (Levitin & Rogers, 2005), which dissociates *pitch memory* from *pitch labeling*, our findings suggest that assessments based on familiar, iconic recordings have construct validity in representing AP memory based on chroma.

Yet, our findings additionally suggest that listeners' judgments cannot be entirely explained in terms of pitch chroma memory. Specifically, we found systematic biases in participant responses, based on the semitone shift of the foil melody, in a manner consistent with the use of *pitch height*. Given the design of the experiment, however, it should be noted that a strategy that solely relied on pitch height would yield at-chance performance (as participants would score 100% when the foil was farther in pitch height and 0% when the foil was closer to the pitch height to the iconic recording). Thus, the pattern of responses can be best described in terms of a hybrid approach, in which both pitch chroma and pitch height contribute to the judgment of familiar melodies.

General discussion

The present study demonstrates that individuals' AP memory for well-known melodies clearly generalizes to novel instances. Participants in all conditions across experiments were significantly above chance in selecting the "correct" version of a familiar melody based on key. Specifically, we saw evidence for generalization across instrumental timbre in both experiments, even when melodies were represented as simplified monophonic recordings. Critically, Experiment 2 demonstrated participants' ability to generalize to a novel octave, meaning judgments of popular melodies cannot be *entirely* explained by using listening strategies related to pitch height. These findings, taken together, suggest that listeners' AP representations for familiar melodies appear to be based on pitch chroma, and thus have similar properties to musical note categories of genuine AP possessors.

These results inform broader models of pitch memory in humans, such as the influential two-step mode of AP, which distinguishes *pitch memory* from *pitch labeling* (Levitin & Rogers, 2005). Pitch memory is considered to be widespread, normally distributed, and implicitly learned through environmental exposure, for example via statistical learning mechanisms (e.g., Saffran et al., 2005). Pitch labeling, in contrast, is a rare ability thought to distinguish genuine AP possessors from non-AP possessors, and the nature of its distribution among the general population (e.g., as a dichotomous or continuously distributed ability) is actively debated. For example, Athos et al. (2007) makes a strong claim that AP is dichotomous, whereas other studies (Bermudez & Zatorre, 2009; Van Hedger et al., 2020) provide evidence that AP is a more continuously distributed ability.

From this two-step model, it is reasonable to predict that only pitch-labeling abilities would afford clear octave and timbre generalization, as pitch memory is tied to specific experiences (e.g., experience with specific recordings) in a way that applying an abstract label to a pitched sound is not. If one is able to apply an abstract category like a musical key to a piece of music, as is the case for an AP possessor, then this ability should manifest regardless of the piece's surface-level features, such as instrumentation or pitch height. In contrast, if listeners have to rely on implicit memories, derived from specific experiences with recordings, then it would be reasonable to expect that listeners would show poor generalization beyond these specific experiences, conceptually similar to the limited generalization (*specificity of learning*) that has been found in other instances of perceptual learning (e.g., Ahissar et al., 2009; Dale et al., 2021).

However, the present findings challenge this view and support the idea that generalization of pitch memory to novel experiences is a widespread ability. Put another way, the findings suggest that listeners are forming more abstract categories that are based on pitch chroma for these well-known melodies. Although this finding is novel in the context of using popular melodies, it can be integrated within a growing body of research on the nature of implicit pitch memory representations. For example, non-AP listeners can tell when isolated notes are played in standard Western tuning (A4 = 440 Hz) compared to notes that have been shifted by 50 cents (Van Hedger et al., 2017), suggesting that listeners have implicitly developed goodness-of-fit representations for individual notes based on listening experience. Furthermore, pitch chroma and musical key has been shown to influence aesthetic evaluations (Ben-Haim et al., 2014) and judgments of musical tension (Eitan et al., 2017) for novel stimuli among non-AP possessors. These findings, taken together, suggest that most listeners have developed robust implicit memories for sounds based on AP and, in particular, pitch chroma.

Yet, there is also evidence that listeners are sometimes influenced by other cues that do not affect the accuracy of AP possessors. For instance, the present study found an interaction between the octave shift and susceptibility to incorrect versions of melodies that were closer in AP height to the iconic recording (Experiment 2, Fig. 3).

Specifically, when participants in the higher octave condition were presented with a trial in which the incorrect melody was shifted one semitone *down*, accuracy was significantly reduced compared to a trial in which the incorrect melody was shifted one semitone shift *up*. The opposite was true in the low octave condition: when the incorrect melody was shifted *up* by one semitone, performance decreased relative to when the incorrect melody was shifted *down* by one semitone. One plausible interpretation of this pattern of results is that participants were influenced by pitch height cues in the experimental paradigm. However, it's important to note that participants were not completely relying on a pitch height strategy; if this were the case, participants would be at chance (50%) given the construction of the experimental design. Thus, the above-chance performance observed suggests a hybrid use of strategies – i.e., that listeners were influenced by both pitch chroma and pitch height. Evidence suggesting the use of a combination of strategies, including both pitch chroma and pitch height, during pitch-related assessments is supported by prior literature; for example, Van Hedger et al. (2015) found that AP possessors were slower at making speeded pitch judgments across multiple octaves, indicating that pitch chroma and pitch height might be integrally linked in memory representations.

There are some limitations inherent in the current study. Given that Experiment 2 was delivered entirely in an online setting, we could not ensure the same level of control over the auditory environment or participant engagement compared to Experiment 1. However, we do not suspect that an online delivery for the second experiment significantly altered or detracted from our findings, for two reasons. First, we incorporated a variety of quality control checks into the online experiment. Notably, participants were pre-screened to ensure that they had met minimum performance criteria in their participation in other MTurk assignments, and participants also were required to pass attention checks throughout the experiment to be included in the main analyses. Second, a robust body of literature has been able to replicate classic auditory findings in an online setting, indicating that when executed properly, online experiments yield equally valid results – with the added benefit of being able to recruit larger, more diverse sample sizes (e.g., see Eerola et al., 2021; Zhao et al., 2022).

A second limitation, which is in part due to Experiment 2 being administered online, is the notably different samples across the two experiments. The sample from Experiment 1 was made up entirely of undergraduate students at the University of Chicago, which confined the sample to a narrow age range (18–24 years). Additionally, the average musical training among the undergraduate sample was high relative to the general population, at an average of 5.87 years with approximately 9 in 10 reporting some musical experience. In contrast, the sample from Experiment 2 featured a much larger age range and lower reported levels of musical training. Because of these differences, it becomes difficult to directly compare performance *across* the two experiments. However, despite these differences in samples, we still observed robust above-chance performance across both experiments for both groups. The data therefore suggest that variability in both musical training and age may not have a strong influence on performance with respect to the given tasks. Although some associations between pitch memory and age have been reported in prior research (Trehub et al., 2008), this was found among children of a comparatively narrow age range (5- to 10-year-olds), and additional research did not find an association between pitch memory and age within overlapping age ranges (9- to 12-year-olds; Schellenberg & Trehub, 2008; 4- to 12-year-olds; Jakubowski et al., 2017). Several additional studies have supported the claim that musical training and age appear to be unrelated to pitch memory for well-known melodies among adults (Jakubowski & Müllensiefen, 2013; Schellenberg & Trehub, 2003; Van Hedger et al., 2018).

Although performance across both experiments was robustly above chance, it is notable that we observed such a large range of individual variability with respect to pitch memory accuracy. Across both experiments, some participants were at (or even below) chance, whereas other participants were close to perfect. Large degrees of individual variability are not inherently unusual in the context of pitch memory paradigms (e.g., see Schellenberg & Trehub, 2003); however, it is intriguing that none of our collected measures of individual differences showed any correlation with pitch memory performance. Despite observing variability along our individual difference measures (years of musical training, active musical practice, and the frequency with which participants listened to pop music), the data suggest that these do not correlate with pitch memory performance in the present context. Some of these nonsignificant associations are perhaps not surprising when considering that prior research has also reported no significant association between factors such as musical training and pitch memory (e.g., Van Hedger et al., 2018). However, the inability to meaningfully explain the large individual differences in the task raise the question of which individual difference measures, if any, might be informative in explaining variation in pitch memory. Based on prior research, one promising approach would be to focus on *performance-based* rather than self-report measures of musical abilities. Specifically, associations were found between pop song pitch memory and singing accuracy (Halpern & Pfordresher, 2022), melodic comparison tasks (i.e., identifying whether two short melodies were the same or different; Jakubowski & Müllensiefen, 2013), and pitch adjustment tasks testing tonal working memory precision (Van Hedger et al., 2018). Given the large individual differences observed in the present experiments, future work might consider both assessing whether these individual differences are stable across time, and examining further how these individual differences in performance relate to performance-based measures of auditory processing, such as those cited here.

Although the present findings are situated within the domain of music cognition, they can also be integrated into a broader learning and memory framework. The finding that most listeners form memory representations for pitch chroma without explicit intention, based on regularities within their environmental input, speaks to the importance of implicit learning mechanisms in underlying complex perceptual representations. The process of implicitly extracting patterns from the environment (e.g., as described in statistical learning) is likely subserving these chroma-based memory representations (i.e., consistently hearing a particular melody in a particular musical key), and has been found to facilitate the formation of perceptual representations in a number of domains beyond music (e.g., see Krogh et al., 2013, and Saffran & Kirkham, 2018, for reviews). Moreover, the observation that listeners have formed memory representations based on pitch chroma – a feature that is putatively inconsequential for melodic identification (e.g., Dowling & Bartlett, 1981; though see Kleinsmith & Neill, 2018) – suggests that "surface-level" perceptual attributes of episodic experiences are not entirely discarded in service of forming memory representations in music. These findings can thus comment on the broader literature examining the extent to which conceptual representations are grounded in the perceptual details of episodic experiences (e.g., Barsalou, 1999; Goldinger, 1998). The fact that listeners were able to generalize across some "surface-level" features (e.g., instrumental timbre, octave) while maintaining an understanding of pitch chroma supports the idea that memory representations for familiar melodies are multidimensional, and the specific weighting of different cues likely depends on several factors including the current task demands (e.g., Saffran et al., 2005).

Overall, the present study finds that everyday listeners not specifically recruited for musical training have robust pitch memory for popular melodies that generalizes across both timbre and octave, suggesting that implicit AP representations developed by the general population are based on pitch chroma and representationally similar to those found in genuine AP. The present study identifies a clear connection between the processes of pitch memory and pitch labeling by demonstrating that the widespread ability of pitch memory is based on representations that are similar to those found among AP possessors. This investigation, as well as future work, can aid our understanding of the nature of AP, as well as the generalizability of human pitch memory more broadly.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13421-023-01429-z.

Author note All data and materials are available via the Open Science Framework at https://doi.org/10.17605/OSF.IO/RYUJE. Experiment 2 was pre-registered through the Open Science Framework (https://doi.org/10.17605/OSF.IO/V3ZBK).

The authors declare no conflicts of interest.

References

Ahissar, M., Nahum, M., Nelken, I., & Hochstein, S. (2009). Reverse hierarchies and sensory learning. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1515), 285–299. https://doi.org/10.1098/rstb.2008.0253

Athos, E. A., Levinson, B., Kistler, A., Zemansky, J., Bostrom, A., Freimer, N., & Gitschier, J. (2007). Dichotomy and perceptual distortions in absolute pitch ability. *Proceedings of the National Academy of Sciences, 104*(37), 14795–14800. https://doi.org/10.1073/pnas.0703868104

Audacity Team (2021). Audacity(R): Free audio editor and recorder [Computer application]. Version 3.0.0. https://audacityteam.org/.

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

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*(4), 577–660. https://doi.org/10.1017/S0140525X99002149

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

Ben-Haim, M. S., Eitan, Z., & Chajut, E. (2014). Pitch memory and exposure effects. *Journal of Experimental Psychology: Human Perception and Performance, 40*(1), 24–32. https://doi.org/10.1037/a0033583

Bermudez, P., & Zatorre, R. J. (2009). A distribution of absolute pitch ability as revealed by computerized testing. *Music Perception, 27*(2), 89–101. https://doi.org/10.1525/mp.2009.27.2.89

Bongiovanni, N. R., Heald, S. L. M., Nusbaum, H. C. & Van Hedger, S. C. (2023). Generalizing across tonal context, timbre, and octave in rapid absolute pitch training. Attention, *Perception, & Psychophysics, 85*(2), 525–542. https://doi.org/10.3758/s13414-023-02653-0.

Dale, G., Cochrane, A., & Green, C. S. (2021). Individual difference predictors of learning and generalization in perceptual learning. *Attention, Perception, & Psychophysics, 83*(5), 2241–2255. https://doi.org/10.3758/s13414-021-02268-3

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

Dowling, W. J., & Bartlett, J. C. (1981). The importance of interval information in long-term memory for melodies. *Psychomusicology: A Journal of Research in Music Cognition, 1*(1), 30–49. https://doi.org/10.1037/h0094275

Eerola, T., Armitage, J., Lavan, N., & Knight, S. (2021). Online data collection in auditory perception and cognition research: Recruitment, testing, data quality and ethical considerations. *Auditory Perception & Cognition, 4*(3–4), 251–280. https://doi.org/10.1080/25742442.2021.2007718

Eitan, Z., Ben-Haim, M. S., & Margulis, E. H. (2017). Implicit absolute pitch representation affects basic tonal perception. *Music Perception, 34*(5), 569–584. https://doi.org/10.1525/mp.2017.34.5.569

Frieler, K., Fischinger, T., Schlemmer, K., Lothwesen, K., Jakubowski, K., & Müllensiefen, D. (2013). Absolute memory for pitch: A comparative replication of Levitin's 1994 study in six European labs. *Musicae Scientiae, 17*(3), 334–349. https://doi.org/10.1177/1029864913493802

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 251–279. https://doi.org/10.1037/0033-295X.105.2.251

Greasley, A. E., & Lamont, A. (2011). Exploring engagement with music in everyday life using experience sampling methodology. *Musicae Scientiae, 15*(1), 45–71. https://doi.org/10.1177/1029864910393417

Halpern, A. R. (1989). Memory for the absolute pitch of familiar songs. *Memory & Cognition, 17*(5), 572–581. https://doi.org/10.3758/BF03197080

Halpern, A. R., & Pfordresher, P. Q. (2022). What do less accurate singers remember? Pitch-matching ability and long-term memory for music. *Attention, Perception, & Psychophysics, 84*(1), 260–269. https://doi.org/10.3758/s13414-021-02391-1

Jakubowski, K., & Müllensiefen, D. (2013). The influence of music-elicited emotions and relative pitch on absolute pitch memory for familiar melodies. *Quarterly Journal of Experimental Psychology, 66*(7), 1259–1267. https://doi.org/10.1080/17470218.2013.803136

Jakubowski, K., Müllensiefen, D., & Stewart, L. (2017). A developmental study of latent absolute pitch memory. *Quarterly Journal of Experimental Psychology, 70*(3), 434–443. https://doi.org/10.1080/17470218.2015.1131726

Juslin, P. N., Liljeström, S., Västfjäll, D., Barradas, G., & Silva, A. (2008). An experience sampling study of emotional reactions to music: Listener, music, and situation. *Emotion, 8*(5), 668–683. https://doi.org/10.1037/a0013505

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

Kleinsmith, A. L., & Neill, W. T. (2018). Recognition of transposed melodies: Effects of pitch distance and harmonic distance. *Psychonomic Bulletin & Review, 25*(5), 1855–1860. https://doi.org/10.3758/s13423-017-1406-5

Krogh, L., Vlach, H. A., & Johnson, S. P. (2013). Statistical learning across development: Flexible yet constrained. *Frontiers in Psychology, 3*. https://doi.org/10.3389/fpsyg.2012.00598

Levitin, D. J. (1994). Absolute memory for musical pitch: Evidence from the production of learned melodies. *Perception & Psychophysics, 56*(4), 414–423. https://doi.org/10.3758/BF03206733

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

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

Marsman, M., & Wagenmakers, E.-J. (2017). Bayesian benefits with JASP. *European Journal of Developmental Psychology, 14*(5), 545–555. https://doi.org/10.1080/17405629.2016.1259614

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

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

Plantinga, J., & Trainor, L. J. (2005). Memory for melody: Infants use a relative pitch code. *Cognition, 98*(1), 1–11. https://doi.org/10.1016/j.cognition.2004.09.008

Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology, 69*(1), 181–203. https://doi.org/10.1146/annurev-psych-122216-011805

Saffran, J. R., Reeck, K., Niebuhr, A., & Wilson, D. (2005). Changing the tune: The structure of the input affects infants' use of absolute and relative pitch. *Developmental Science, 8*(1), 1–7. https://doi.org/10.1111/j.1467-7687.2005.00387.x

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

Schellenberg, E. G., & Trehub, S. E. (2008). Is there an Asian advantage for pitch memory? *Music Perception, 25*(3), 241–252. https://doi.org/10.1525/mp.2008.25.3.241

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

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

Terhardt, E., & Seewann, M. (1983). Aural key identification and its relationship to absolute pitch. *Music Perception, 1*(1), 63–83. https://doi.org/10.2307/40285250

Trehub, S. E., Glenn Schellenberg, E., & Nakata, T. (2008). Cross-cultural perspectives on pitch memory. *Journal of Experimental Child Psychology, 100*(1), 40–52. https://doi.org/10.1016/j.jecp.2008.01.007

Van Hedger, S. C., Heald, S. L. M., Huang, A., Rutstein, B., & Nusbaum, H. C. (2017). Telling in-tune from out-of-tune: Widespread evidence for implicit absolute intonation. *Psychonomic Bulletin & Review, 24*(2), 481–488. https://doi.org/10.3758/s13423-016-1099-1

Van Hedger, S. C., Heald, S. L. M., & Nusbaum, H. C. (2015). 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. (2018). 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., 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

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

Zhao, S., Brown, C. A., Holt, L. L., & Dick, F. (2022). Robust and efficient online auditory psychophysics. *Trends in Hearing, 26*, 233121652211187. https://doi.org/10.1177/23312165221118792

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.