Cognition 154 (2016) 139–150

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Original Articles

What the [bleep]? Enhanced absolute pitch memory for a 1000 Hz sine tone

Stephen C. Van Hedger *, Shannon L.M. Heald, Howard C. Nusbaum

Department of Psychology, The University of Chicago, Chicago, IL 60637, USA

ARTICLE INFO

Article history:
Received 13 October 2015
Revised 11 March 2016
Accepted 1 June 2016

Keywords:
Absolute pitch
Auditory memory
Expertise
Memory precision

ABSTRACT

Many individuals are able to perceive when the tuning of familiar stimuli, such as popular music recordings, has been altered. This suggests a kind of ubiquitous pitch memory, though it is unclear how this ability differs across individuals with and without absolute pitch (AP) and whether it plays any role in AP. In the present study, we take advantage of a salient single frequency – the 1000 Hz sine tone used to censor taboo words in broadcast media – to assess the nature of this kind of pitch memory across individuals with and without AP. We show that non-AP participants are accurate at selecting the correct version of the censor tone among incorrect versions shifted by either one or two semitones, though their accuracy was still below that of an AP population (Experiment 1). This suggests a benefit for AP listeners that could be due to the use of explicit note categories or greater amounts of musical training. However, AP possessors still outperformed all non-AP participants when incorrect versions of the censor tone were shifted within a note category, even when controlling for musical experience (Experiment 2). Experiment 3 demonstrated that AP listeners did not appear to possess a category label for the censor tone that could have helped them differentiate the censor tones used in Experiment 2. Overall, these results suggest that AP possessors may have better pitch memory, even when divorced from pitch labeling (note categories). As such, these results have implications for how AP may develop and be maintained.

1. Introduction

Absolute pitch (AP) is defined as the ability to explicitly name or produce a musical note without the aid of a reference note (e.g., Ward, 1999). Despite the consistency in how AP is defined, the rarity of AP, along with the notion that AP is dichotomous (one either possesses AP or does not), have been central points of debate over the past century of research. In terms of occurrence, AP is often cited as manifesting in every one in 10,000 individuals (Bachem, 1955; Profita & Bidder, 1988), though this estimate does not have strong empirical support, and there are likely several important factors in determining the true rate at which AP occurs. For example, there appear to be large cultural differences in the occurrence of AP. Miyazaki, Makomaska, and Rakowski (2012) reported that 30% of Japanese music students possessed "true" AP, whereas only 7% of Polish music students possessed "true" AP. Similar differences have also been reported in AP prevalence between students at music conservatories in the United States versus China, with the latter group demonstrating superior absolute pitch performance (e.g., Deutsch, Henthorn, Marvin, & Xu, 2006), likely due to differential early experience with a tonal language (e.g., Deutsch, Dooley, Henthorn, & Head, 2009).

Moreover, the use of terms like "true AP" highlights the fact that some individuals may display AP-like ability, even if they are not able to identify or produce musical notes with sufficient speed and accuracy as to be classified as a "true" AP possessor. Despite the variability in performance that exists both within an "AP population" and a "non-AP population," AP is still often discussed in dichotomous terms (e.g., Athos et al., 2007; though see Bachem, 1937; as well as Bermudez & Zatorre, 2009). Thus, while recent research has begun to reevaluate the rarity and dichotomy of AP, these are still common terms used to describe the ability.

Despite the putative rarity and dichotomy of the ability to explicitly name or produce an isolated musical note, an increasing amount of research supports the idea that many people have some absolute pitch memory, even if they cannot explicitly label an isolated pitch with its musical note name. This more widespread pitch memory allows individuals to correctly identify or produce the correct absolute key of a familiar song (Halpern, 1989; Jakubowski & Müllensiefen, 2013; Levitin, 1994; Schellenberg & Trehub, 2003, 2008; Terhardt & Seewann, 1983), and to identify a correctly-pitched version of certain non-musical items, such as a landline dial tone (Smith & Schmuckler, 2008). In some extreme circumstances,

---

* Corresponding author at: 5848 S. University Ave. B406, Chicago, IL 60637, USA.
E-mail address: shedger@uchicago.edu (S.C. Van Hedger).

http://dx.doi.org/10.1016/j.cognition.2016.06.001
0010-0277/© 2016 Elsevier B.V. All rights reserved.

this pitch memory may even allow an individual to remember and reproduce a single pitch – after hearing a number of interfering tones – even if they cannot explicitly label the to-be-remembered pitch, which is generally thought to not be possible without the aid of an explicit label (Ross, Olson, & Gore, 2003).

This general kind of pitch memory, which does not require the explicit categorization or labeling of pitches, is sometimes described as the first step of the proposed two-step process underlying true AP ability, with the second step being the ability to apply an explicit musical note label to pitch information (Levitin & Rogers, 2005). Evidence of widespread pitch memory has led researchers to suggest that the general ability to form long-term pitch memories might be normally distributed in the population, with true AP only being differentiated by the explicit ability to apply a long-term note category to pitches (Schellenberg & Trehub, 2003). Moreover, while it has been assumed that this more general pitch memory necessarily requires extensive experience with hearing stimuli at the same pitch level (e.g., through hearing a particular music recording several dozen times), more recent research has suggested that this kind of pitch memory can be reliably established even after a single exposure (Schellenberg & Habashi, 2015; Schellenberg, Stalinski, & Marks, 2014)

If true AP is distinguished through the explicit knowledge of note categories, then it is possible that AP possessors might not show any enhancements in general pitch processing or pitch memory precision compared to non-AP possessors. In line with this reasoning, previous research has found that AP possessors are not particularly "gifted" when it comes to basic auditory processing abilities. AP possessors do not have an enhanced ability to resolve frequency, spatial, or temporal differences in sounds (Fujisaki & Kashino, 2002), suggesting that low-level differences in perceptual discrimination are not likely related to AP. Moreover, while AP possessors appear to have better long-term memory for pitch compared to non-AP possessors (e.g., Rakowski & Morawska-Bungeler, 1987), it has been suggested that this is *not* because AP possessors are better at remembering the "sound of a tone," but rather because they can identify the tone by its note name and remember this in long-term memory (Takeuchi & Hulse, 1993, p. 354). This argument is supported by a number of empirical observations. For example, in a task where participants had to judge which of two tones was higher after varying delays between tones, Siegel (1974) found no difference between AP and non-AP possessors, even after long retention intervals, when the difference between the two tones was within a note category (e.g., if both tones would be classified as "C"). This suggests that AP possessors outperform non-AP possessors on such pitch memory tasks because they can remember a category label, not because they remember the fine-grained details of the pitch. In production, AP possessors are biased in their reproduction of mistuned pitches, such that they are more likely to produce a pitch that conforms to an in-tune note, particularly with longer intervals between hearing the note and producing the note (Hutchins, Hutka, & Moreno, 2015). To be clear, these kinds of results suggest that there is no difference in auditory sensory processing for AP and non-AP possessors, and further, there is no difference in long-term pitch memory between AP and non-AP possessors, at least when the pitches cannot be differentiated at the note category level. The difference between these groups is presumably in the knowledge of the note labels that correspond to musical pitches.

Therefore, if AP is differentiated from non-AP not through enhancements in auditory processing or perceptual memory, but rather through the availability of explicit category labels, then one might predict that more general measures of pitch memory (e.g., identifying a correctly tuned version of the theme to "The X-Files") might not differ between AP and non-AP possessors, controlling for any possible strategic use of explicit musical note knowledge (e.g., prior knowledge that "The X-Files" theme is in A minor). Controlling for the use of explicit musical note knowledge among an AP population, however, is not trivial. Dooley (2011) – the only study assessing these kinds of pitch memory differences between AP and non-AP possessors using familiar musical stimuli – tried to control for the use of explicit category knowledge by limiting participants to reproducing music pieces for which they did not see sheet music or explicitly check against an instrument. However, it could be argued that an AP population would still potentially have explicit note knowledge that could be used to help them deduce whether a previously heard piece was in the correct key, regardless of whether they had previously played the piece or seen sheet music for the piece. This potential confound makes it unclear whether long-term pitch memory is truly better in an AP population. While some case studies have supported the idea that true AP may be grounded in a fundamentally different (and superior) means of absolutely encoding pitch that is independent of musical note labeling (Ross, Gore, & Marks, 2005; Ross & Marks, 2009; Ross et al., 2003), it is unclear from these cases whether a long-term pitch memory for a well-known stimulus (e.g., a familiar music recording) would similarly differ between AP and non-AP groups when explicit labels are not beneficial for performance.

The present study provides a novel means for assessing the nature of pitch memory across both AP and non-AP possessors by taking advantage of a particularly salient frequency – the 1000 Hz sine tone used to censor taboo words in broadcast media. Using the censor tone marks an important deviation from previous pitch memory studies in several ways. Most notably, this frequency does not correspond to a correctly tuned musical note, falling between the notes B5 and C6. As such, it might inherently challenge the explicit category labels that an AP possessor might use in a test of pitch memory (cf. Rakowski, 1972). In the current set of experiments, we specifically address whether, in this particular situation, pitch memory accuracy will be comparable across AP and non-AP populations, or whether AP possessors will show enhanced absolute memory for the censor tone compared to non-AP possessors.

2. Experiment 1

2.1. Methods

2.1.1. Participants

473 individuals participated in Experiment 1. There were three total participant groups. The first group, hereafter referred to as the "MT1" group ($n = 200$), was recruited through Amazon's Mechanical Turk (mTurk). The second group, hereafter referred to as the "MT2" group ($n = 200$), was also recruited through Amazon's mTurk, though they listened to a modified audio clip (see Section 2.1.2 for details). The third group, hereafter referred to as the "AP" group ($n = 73$), consisted of self-identified AP possessors, who completed the same procedure as the MT1 group.

All participants completed the study online through Qualtrics (Qualtrics: Provo, UT). All Mechanical Turk participants (MT1 and MT2 groups, total $n = 400$) had to be residing in the United States and had to have completed at least 50 prior mTurk assignments with an approval rating of 90% or higher.

2.1.2. Materials and procedure

The five sine tones, which served as the different versions of the censor tone, were generated in Adobe Audition (Adobe Systems: San Jose, CA). The correct version was generated at 1000 Hz. The sharp versions were generated at 1059.46 Hz (one semitone) and 1122.46 Hz (two semitones), while the flat versions were generated at 943.87 Hz (one semitone) and 890.90 Hz (two semitones). For reference, one semitone corresponds to approximately a 5.9% pitch change and is the smallest pitch difference used in conventional Western music. The primary audio clip, which was presented to the MT1 and AP groups, was 22.5 s in duration and was taken from an uncensored George Carlin comedy routine, in which George Carlin talks about the nature of discovering swear words as a child. George Carlin swears twice over the course of the audio clip (once at approximately 13.5 s, once at approximately 16.5 s). We silenced the audio of the taboo words, filling in the silence with one of the generated stimulus tones. Participants always heard the same-pitched stimulus tone within the clip (i.e. they would not hear a different pitch at 13.5 s and 16.5 s). Each participant heard the George Carlin audio clip twice – once with correct censor tone, and once with one of the incorrect stimulus tones. We counterbalanced the presentation order of the correct version and incorrect version across participants, as well as the nature of the incorrect version (one semitone flat, two semitones flat, one semitone sharp, or two semitones sharp) across participants. Thus, there were eight total versions of the experiment to which participants were randomly assigned.

The MT2 group ($n = 200$) completed an alternate version of the task. To ensure that the results from the first group of participants could not be solely attributed to familiarity with George Carlin's voice or the specific comedy routine, the second group of participants heard the stimulus tone in a different context. Participants heard two versions of an AT&T text-to-speech (TTS) synthesized female voice say, "I don't know what the [bleep] you are talking about," with one version at the correct absolute pitch and the other version at an incorrect absolute pitch (using the same counterbalancing as the first group). Each stimulus tone occurred approximately 1.0 s into the sound clip, and the entire recording was approximately 2.7 s long.

All groups were told that they would hear two audio clips containing the "bleep" used to censor inappropriate words, with one of the two versions containing the correctly pitched "bleep" (i.e. the exact same sound that would be heard on TV or on the radio) and the other version containing an incorrectly pitched "bleep" (which would sound "too high" or "too low" compared to the version heard on TV or on the radio). Participants were also told that they would not know whether the correct version would be presented first or second. After hearing both versions of the audio clip, participants made a forced choice judgment with no feedback as to which version they thought contained the correct version of the censor tone. Fig. 1A shows the location of the correct censor tone and incorrect versions used in Experiment 1 relative to Western musical notes.

After making this judgment, all participants were asked to rate their confidence in their answer, as well as their familiarity with the censor tone (both on a 100 point slider scale). Participants were then asked whether they possessed AP (with the response options of *yes*, *no*, and *not sure*). AP was defined to participants as the ability to name or produce any musical note without the aid of a reference note. Finally, participants were asked whether they had previously participated in any study in which they were asked to judge the pitch of the censor tone.

2.1.3. Participant inclusion

If participants reported no familiarity with the censor tone, they were excluded from all analyses. Moreover, any participant who reported participating in a previous study on judging the pitch of the censor tone was also excluded from all analyses. AP possession was used as a criterion for exclusion for all groups with the exception of the AP group, in which AP possession was used as a criterion for inclusion.

In the MT1 group, 40 participants did not unequivocally self-report as not possessing AP (9 participants reported possessing AP and 31 participants were "not sure" whether they possessed AP). Participants in the MT1 group were asked an additional question about the subject matter of the audio clip (choosing between "religion," "baseball," "swearing," and "politics"). If participants did not choose "swearing," then we surmised that participants either did not listen to the sound clips, or had great difficulty understanding the task. Of the 160 participants who explicitly stated that they did not possess AP, 9 participants failed to choose the correct answer. We excluded these participants from our

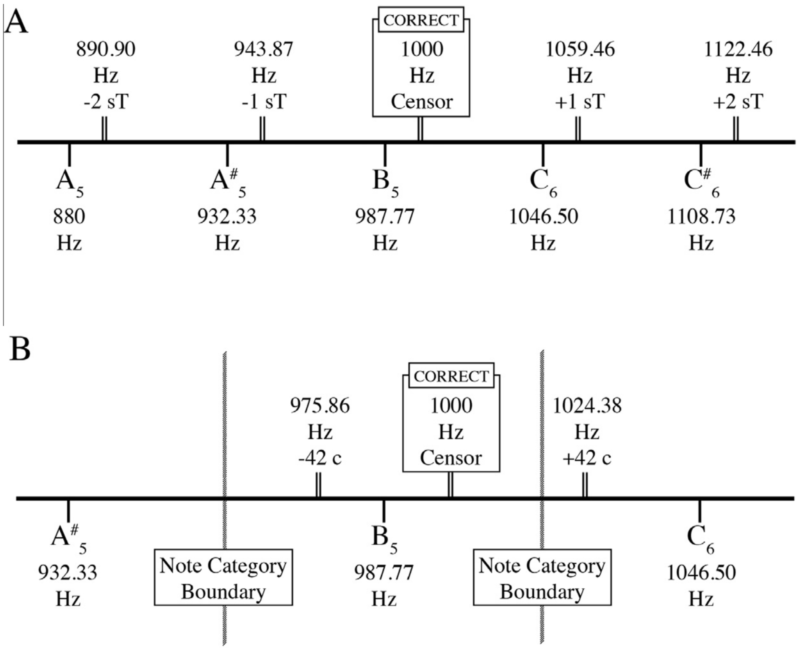

Fig. 1. Location of the different versions of the censor tone relative to perfectly tuned Western musical notes. In Experiment 1 (A), the location of the incorrect versions are exactly one or two semitones removed from the correct censor tone and thus fall 21 cents sharp from Western tunings. In Experiments 2A and 2B (B), the location of the incorrect version are 42 cents removed from the correct version of the censor tone, and thus the flat version and correct version would be classified as the note "B" whereas the sharp version would be classified as the note "C".

analyses, leaving 151 analyzable participants in the MT1 group. 74 of the 151 participants were in the 2 semitone condition, while 77 of the 151 participants were in the 1 semitone condition.

Of the 73 participants recruited in the AP group, 1 participant self-identified as not possessing AP. The remaining participants were given a short test of their AP ability, in which they heard seven notes in rapid succession (E5, Eb2, F#3, G4, Eb6, G4, and D3) and were asked to label the final note. Participants either needed to correctly classify the final note or be within one semitone of the correct answer in order to be included in the analyses. Five participants did not pass this short AP test. Of the remaining participants, 4 reported having no familiarity with the censor tone. We thus analyzed the remaining 63 participants. Of the 63 participants, 31 were in the 2-semitone condition, while 32 were in the 1-semitone condition.

In the MT2 group, 51 participants did not unequivocally self-identify as not possessing AP (13 reported possessing AP and 38 were "not sure" whether they possessed AP). Of the remaining participants, 12 reported having no familiarity with the censor tone. Of the remaining participants, 13 reported previously participating in a censor tone experiment (likely in the MT1 sample, as the MT2 sample was collected later in time). We thus analyzed the remaining 124 participants. 72 of these participants were in the 2-semitone condition, while 52 of these participants were in the 1-semitone condition.

While participant inclusion across these samples is conservative (omitting 24.5% of the MT1 sample, 13.7% of the AP sample, and 38.0% of the MT2 sample), given the hypotheses of the study this conservative inclusion criterion is justified. To additionally assess whether any of our reported results would change as a function of a more liberal inclusion criterion, we modified our participant inclusion criteria to include participants who were "not sure" of their AP ability but passed all other qualifications (previous familiarity with the censor tone, no previous participation in a censor tone experiment, and successfully answering the audio content question for the MT1 group). This more liberal inclusion criterion added 28 participants to the MT1 sample (total $n = 179$), 0 participants to the AP sample (as all participants either definitely reported possessing or not-possessing AP, total $n = 63$), and 30 participants in the MT2 sample (total $n = 154$). A comparison of the results from both the conservative and liberal inclusion criteria is reported in Table 1.

Table 1
Comparison of conservative and liberal participant inclusion from Experiment 1.

Group	Inclusion	Shift	Sum	Total	Proportion
MT1	Liberal	1 Semitone	66	90	0.733
		2 Semitones	76	89	0.854
	Conservative	1 Semitone	55	77	0.714
		2 Semitones	63	74	0.851
MT2	Liberal	1 Semitone	52	71	0.732
		2 Semitones	66	83	0.795
	Conservative	1 Semitone	40	52	0.769
		2 Semitones	59	72	0.819
AP	Liberal	1 Semitone	29	32	0.906
		2 Semitones	30	31	0.968
	Conservative	1 Semitone	29	32	0.906
		2 Semitones	30	31	0.968

2.1.4. Data analysis plan

To assess performance, we used a Bayesian equivalent of a binomial test (Bååth, 2014), in which the relative frequency of success can be estimated using Markov Chain Monte Carlo (MCMC) simulations. We first ran 15,000 MCMC simulations with a non-informed beta prior of (1, 1). We collapsed across the dimensions of flat/sharp as there were no significant differences between these versions in this experiment (Fisher's Exact Test: $p = 0.69$ for MT1, $p = 0.64$ for MT2, $p = 0.35$ for AP). Furthermore, we collapsed across versions in which the correct version was presented first versus second, as we specifically counterbalanced presentation order across participants to mitigate order effects. Thus, we specifically examined accuracy as a function of semitone distance (one semitone removed from the correct version versus two semitones removed from the correct version).

2.2. Results

2.2.1. Non-AP possessors

All groups of participants were well above chance at selecting the correct frequency of the stimulus tone. For the MT1 group, 55 out of 77 (71.4%) participants selected the correct version in the one-semitone condition, and 63 out of 74 (85.1%) participants selected the correct version in the two-semitone conditions. The 95% credible interval was (0.61, 0.81) for the one-semitone conditions and (0.76, 0.92) for the two-semitone conditions. Both the one- and two-semitone conditions had a 0.999 probability that the relative frequency of success was more than 50% (chance).

The second Mechanical Turk sample (MT2) sample – in which listeners selected the correct frequency of the censor tone in a different context of an unknown speaker – nevertheless showed the same pattern as the MT1 sample. When the incorrect version of the censor tone was one semitone removed from the correct version, 40 of 52 (76.9%) participants were able to select the correct version. This number improved to 59 out of 72 (81.9%) when the incorrect version was removed from the correct version by two semitones. The 95% credible intervals were (0.64, 0.87) and (0.72, 0.90) for the one-semitone and two-semitone conditions, respectively, and thus did not overlap with the chance estimate of 0.50.

To compare accuracy in the present experiment to prior research in which participants judged the pitch of popular music recordings, we compared performance in the present experiment to the results of Schellenberg and Trehub (2003). To do so, we recomputed MCMC simulations with a different prior probability. Specifically, we used 58% for the one-semitone conditions and 70% for the two-semitone conditions, as these were the reported mean accuracy levels in Schellenberg and Trehub (2003). For the MT1 group, there was a 0.998 probability that the relative frequency of success was more than 58% for the one-semitone conditions and a 0.999 probability that the relative frequency of success was more than 70% for the two-semitone conditions. For the MT2 group, there was a 0.998 probability that the relative frequency of success was more than 58% for the one-semitone conditions, and a 0.987 probability that the relative frequency of success was more than 70% for the two-semitone conditions.

This performance difference, however, could have been partly due to differences in experimental setup. Specifically, Schellenberg and Trehub (2003) included practice trials, as well as several blocks of pitch judgments during which participants could have developed an interfering pitch memory for the incorrect versions of recordings. We thus compared performance in the present experiment with another prior – that from Schellenberg and Trehub (2008), in which there were no practice trials. Specifically, we only examined performance from the first block of Schellenberg and Trehub (2008), as there was evidence that performance decreased in subsequent blocks, likely due to the development of a memory representation for the incorrect recordings. Accuracy in the first block was around 78%, though this was only using incorrect tunings of two semitones. Comparing the present non-AP performance to this prior, there was a 0.923 probability that the relative frequency of success was greater than 78% for the MT1 sample, and a 0.763 probability that the relative frequency of success was greater than 78% for the MT2 sample.

Combining both samples, there was a 0.949 probability that the relative frequency of success was greater than 78%. Thus, while performance in the present experiment was perhaps not qualitatively different from previous studies on pitch memory, we found strong evidence that pitch memory for the censor tone is at least as strong as pitch memory for well-known music recordings, possibly even representing the upper limits of this kind of pitch memory (i.e., comparable to the most accurate music recordings).

2.2.2. AP possessors

Despite the high performance among all non-AP groups (73.6% accuracy in the one-semitone conditions, averaged across samples and 83.6% accuracy in the two-semitone conditions, averaged across samples), there was strong evidence that the AP possessors performed better than the non-AP possessors. Specifically, 29 of 32 (90.1%) AP possessors chose the correct censor tone in the one-semitone conditions, while 30 of 31 (96.7%) AP possessors chose the correct censor tone in the two-semitone conditions. The 95% credible intervals were (0.77, 0.98) and (0.86, 1.00) for the one- and two-semitone conditions, respectively. Moreover, the probabilities that the AP possessors' frequency of success was higher than the non-AP population were 0.971 and 0.961 for the one- and two- semitone conditions, respectively. The performance across the MT1, MT2, and AP groups is displayed in Fig. 2.

2.3. Discussion

The purpose of Experiment 1 was to assess pitch memory among AP and non-AP possessors for a salient and isolated frequency that is encountered in one's environment – the 1000 Hz censor tone. We did not find consistent evidence that memory for the censor tone was qualitatively different from previous accounts of pitch memory among non-AP possessors (Schellenberg & Trehub, 2003, 2008). That being said, AP possessors were reliably more accurate than non-AP possessors and virtually at ceiling performance for both semitone conditions, which is notable considering the fact that the censor tone – while pitched – does not correspond to an in-tune note, is not generally thought of as musical, and consequently might not have been explicitly categorized by AP possessors prior to the experiment. The difference in performance between AP and non-AP possessors conceptually supports previous research in which AP possessors have been shown to perform better than non-AP possessors in general tests of pitch memory.

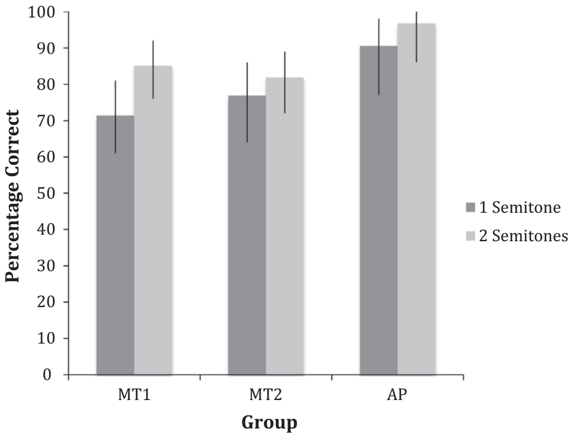

Fig. 2. Accuracy in selecting the correctly tuned censor tone among an alternative that was shifted by one semitone (dark grey) or two semitones (light grey) across the three groups of Experiment 1. All groups were well-above chance in selecting the correct censor tone (corresponding to 50%). Error bars represent 95% credible intervals.

While enhanced performance in the AP population suggests that AP possessors may indeed have better pitch memory than non-AP possessors, one potential concern with the results of Experiment 1 is that AP possessors might have been able to use *both* pitch memory and explicit note knowledge (i.e. knowledge of note labels). While this explanation may seem unlikely, especially given the peculiarities of the censor tone as it relates to the Western musical system, the possibility that AP possessors could rely on explicit labeling knowledge to perform the task (e.g., through knowing that the censor tone roughly corresponds to a "B") cannot be ruled out in the present experiment. Experiment 2 addresses this concern by introducing a modified version of the task, in which the incorrect versions of the censor tone were shifted by only 42 cents, rather than a full semitone (100 cents) or two semitones (200 cents) and thus were within the same note category rather than between categories. By having participants select the correct version of the censor tone among alternative tones that could be classified as belonging to the same note category, a closer comparison of pitch memory between non-AP and AP possessors independent of explicit pitch labeling ability becomes possible (cf. Siegel, 1974).

A secondary concern with comparing AP and non-AP possessors in Experiment 1 has to do with potential overall music experience differences between these groups. Given that active musical instruction can influence difference limens in pitch perception (Kishon-Rabin, Amir, Vexler, & Zaltz, 2001; though see Micheyl, Delhommeau, Perrot, & Oxenham, 2006 for a more rapid account of improving pitch discrimination thresholds), it might be the case that AP possessors are only better than non-AP possessors because they have overall greater amounts of music experience. In line with this reasoning, while significant musical experience does not seem to be *necessary* for long-term pitch memory (cf. Schellenberg & Trehub, 2003), some research has demonstrated that music experience influences the accuracy of pitch memory (e.g., Frieler et al., 2013).

Given these concerns, we addressed two main questions in Experiment 2. First, would we still observe an AP advantage in pitch memory when designing a task in which explicit pitch labeling would not be beneficial? Second, would any AP advantages in pitch memory persist when factoring in overall music experience?

3. Experiment 2A

3.1. Methods

3.1.1. Participants

337 individuals participated in the Experiment 2A. There were two participant groups – a non-AP group ($n = 204$), which was recruited through mTurk using the same qualifications as Experiment 1, and an AP group ($n = 133$).

3.1.2. Materials and procedure

The three sine tones that served as the different versions of the censor tone were generated in Adobe Audition (Adobe Systems: San Jose, CA). The correct version was generated at 1000 Hz. The sharp version was generated at 1024.38 Hz (42 cents above 1000 Hz), while the flat version was generated at 975.86 Hz (42 cents below 1000 Hz). The reason we chose 42 cents as the magnitude of difference between the correct and incorrect censor tones has to do with the location of 1000 Hz relative to the standard note categories of Western music. Since 1000 Hz corresponds to a B5 that is 21 cents sharp, the flat version of the censor tone is essentially mirrored around a perfectly in-tune B5 (as it is 21 cents flat) and thus is still well-within the category of "B". On the other hand, the sharp version of the censor tone – being 42 cents higher than

1000 Hz, is actually closer to C6, as it is 63 cents sharp relative to a perfectly tuned B5 (compared to 37 cents flat relative to a perfectly tuned C6). Fig. 1B displays this tuning difference. Thus, while both incorrect censor tones were equally spaced and less than a semitone removed from the correct censor tone, we hypothesized that AP possessors might show improved performance for the sharp condition, as this incorrect version traversed the putative note category boundary (thus making explicit note knowledge a potentially useful piece of information). Non-AP possessors, on the other hand, were hypothesized to show no asymmetry in performance as they do not possess explicit note knowledge.

The procedure was identical to that used in the MT2 sample from Experiment 1 (using the synthesized talker rather than the comedy routine from George Carlin). As in Experiment 1, participants were asked to rate their familiarity with the censor tone and were asked if they possessed AP. Participants were additionally asked whether they had participated in any previous study in which the pitch of the censor tone was judged.

All participants were asked to provide an estimate of musical instruction after making their selection. There were five answer options, corresponding to five graded levels of music experience. Music experience was defined as the number of years an individual had actively played a musical instrument, including the voice. The first option corresponded to no musical instruction. The second option corresponded to 1–3 years of instruction. The third option corresponded to 4–6 years of instruction. The fourth option corresponded to 7–9 years of instruction, and the fifth option corresponded to more than 9 years of instruction.

3.1.3. Participant inclusion

Of the 133 participants recruited in the AP group, 46 did not unequivocally self-identify as possessing AP (10 reported not possessing AP, 35 were “not sure” of their AP ability, and 1 failed to answer the question). Of the remaining participants, 8 were not able to pass the short test of AP ability (which was identical to the one used in Experiment 1). Of the remaining participants, 16 reported that they had participated in Experiment 1 and one participant reported no familiarity with the censor tone. This left 62 participants in our analyses.

Of the 204 participants recruited in the non-AP group, 48 did not unequivocally self-reported as not possessing AP (9 self-reported as possessing AP and 39 were “not sure” of their AP ability). Of the remaining participants, 10 reported that they had participated in Experiment 1. Of the remaining participants, 8 reported not having any familiarity with the censor tone. We omitted these participants from all analyses, thus leaving 138 participants in our analyses.

While participant inclusion may seem conservative, similar to Experiment 1 we assessed whether our results significantly changed as a function of a more liberal inclusion criterion. Specifically, we included individuals in the AP group who were “not sure” of their AP ability but nonetheless passed the short test for AP ability (total added $n = 26$), and we included the individuals in the non-AP group who were “not sure” of their AP ability (total added $n = 32$). These added participants had to pass all other measures for inclusion (reported familiarity with the censor tone, no reported participation in a previous censor tone study). Adding these participants to our analyses did not significantly change any of our conclusions. A comparison of the results from using conservative and liberal inclusion criteria is reported in Table 2.

Table 2
Comparison of conservative and liberal participant inclusion from Experiments 2A (Non-AP, AP) and 2B (Expert).

Group	Inclusion	Shift	Sum	Total	Proportion
Non-AP	Liberal	Flat	55	88	0.625
		Sharp	53	82	0.646
	Conservative	Flat	43	72	0.597
		Sharp	42	66	0.636
AP	Liberal	Flat	40	49	0.816
		Sharp	32	39	0.821
	Conservative	Flat	27	34	0.794
		Sharp	24	28	0.857
Expert	Liberal	Flat	126	202	0.623
		Sharp	149	208	0.716
	Conservative	Flat	117	189	0.619
		Sharp	145	202	0.718

3.2. Results

Despite the more difficult nature of the task due to the reduced frequency difference between the censor tones, performance in the non-AP group was still reliably above chance, at least when collapsed across conditions. When the incorrect version was 42 cents flat, 43 out of 72 (59.7%) participants chose the correct version of the censor tone. The 95% credible interval was (0.48, 0.70), and there was a 0.949 probability that the relative frequency of success was more than 50% (chance). When the incorrect version was 42 cents sharp, 42 out of 66 (63.6%) participants chose the correct version of the censor tone. The 95% credible interval was (0.52, 0.75), and there was a 0.987 probability that the relative frequency of success was greater than chance. Overall, 85 of 138 (61.6%) non-AP participants chose the correct censor tone. The 95% credible interval was (0.54, 0.70) and there was a 0.997 probability that the relative frequency of success was greater than chance.

Performance in the AP group, however, was still better relative to the non-AP group. When the incorrect version was 42 cents flat, 27 out of 34 (79.4%) participants chose the correct version of the censor tone. The 95% credible interval was (0.64, 0.90), and there was a 0.991 probability that the relative frequency of success was greater than the performance observed in the non-AP group. When the incorrect version was 42 cents sharp, 24 out of 28 (85.7%) participants chose the correct version of the censor tone. The 95% credible interval was (0.70, 0.95), and there was a 0.993 probability that the relative frequency of success was greater than the performance observed in the non-AP group.

To address the possibility that perhaps AP possessors were performing better than non-AP possessors due to differences in musical experience, we looked at group differences when controlling for overall music experience by constructing a generalized linear model with a binomial link. Music experience (treated as a single, continuous variable with 5 levels) and group (AP versus non-AP) were our predictor variables, and accuracy was our dependent variable. Even when controlling for music experience, the difference in accuracy between the AP and non-AP groups was significant ($\beta = 1.27$, $SE = 0.48$, $p = 0.008$), suggesting that AP possessors have a better pitch memory for the censor tone even when controlling for music experience. Adding self-reported familiarity with the censor tone to the model also did not change the significant difference between the AP and non-AP groups ($\beta = 1.37$, $SE = 0.49$, $p = 0.005$). Music experience was not a significant predictor of censor tone accuracy ($\beta = -0.10$, $SE = 0.13$, $p = 0.43$).

4. Experiment 2B

4.1. Methods

4.1.1. Participants

To further assess whether performance differences between our AP and non-AP groups could not be attributed to differences in music experience, we recruited an additional group of self-identified musicians, hereafter referred to as the "Expert" group ($n = 456$). Fig. 3 displays the histogram of self-reported musical experience across the non-AP and AP groups of Experiment 2A, as well as the Expert group of Experiment 2B. The number of participants in each music training bin across the AP and Expert groups was not significantly different [X^2 (4) = 7.30, $p = 0.12$].

4.1.2. Materials and procedure

The materials and procedure were identical to those used in Experiment 2A.

4.1.3. Participant inclusion

Of the 456 total participants, 55 participants did not explicitly report being non-AP possessors (15 explicitly self-reported as possessing AP, 39 reported being "not sure" of whether they possessed AP, and 1 left the question blank). Of the 15 who reported possessing AP, 12 passed the check for AP ability used in Experiments 1 and 2A, and were thus included in the AP group analyses of Experiment 2A. Of the remaining participants, 6 reported previously participating in a study in which they were asked to judge the pitch of the censor tone. Of the remaining participants, 4 reported having no familiarity with the censor tone. These participants were thus excluded, leaving 391 participants in the final analyses.

Similar to Experiments 1 and 2A, we adopted a more liberal criterion for participant inclusion to assess whether this would change the nature of our results. Specifically, we added 19 participants who reported that they were "not sure" of their AP ability and did not pass the AP check (i.e. they were not within one semitone of the correct answer). There were thus 410 analyzable participants in the liberal analyses. The comparison of the conservative and liberal inclusion criterion for all groups across Experiment 2A and 2B are reported in Table 2.

4.2. Results

Despite the high level of reported musical experience within the Expert group, overall performance for selecting the correctly tuned censor tone appeared to fall in-between the non-AP (non-Expert) and AP groups of Experiment 2A. When the incorrect version was 42 cents flat, 117 out of 189 (61.9%) participants chose the correct version of the censor tone. The 95% credible interval was (0.55, 0.69), and there was a 0.999 probability that the relative frequency of success was *less* than the level of performance seen among AP possessors in Experiment 2A. When the incorrect version was 42 cents sharp, 145 out of 202 (71.8%) participants chose the correct version of the censor tone. The 95% credible interval was (0.65, 0.78), and there was a 0.999 probability that the relative frequency of success was *less* than the level of performance seen among AP possessors in Experiment 2A. Unexpectedly, the Expert group showed an asymmetry in performance between flat and sharp conditions that was hypothesized to only exist in the AP group. Specifically, Experts were approximately 9.9% more accurate in the sharp conditions compared to flat conditions, which was a significant difference (Fisher's Exact Test: $p = 0.04$).

There was some evidence that the Expert group performed better than the non-AP group from Experiment 2A, lending support to the idea that music training might help pitch memory performance to an extent. There was a 0.73 probability that the performance among the Expert group was higher than the non-AP group for the flat conditions and a 0.993 probability that the performance among the Expert group was higher than the non-AP group for the sharp conditions. Overall (collapsing across flat and sharp conditions), there was a 0.985 probability that performance in the Expert group was higher than performance in the non-AP group Despite technically falling between the accuracy levels of the non-AP and AP groups of Experiment 2A, it is clear that the Expert group was much more comparable to the non-AP group as opposed to the AP group. Specifically, while the Expert group performed about 2 points higher on the flat conditions and about 8 points higher on the sharp conditions than the non-AP group, this level of performance was still about 17 points lower on the flat conditions and 14 points lower on the sharp conditions than the AP group. Fig. 4 displays how the performance from the Expert group in the current experiment compares to the performance from the non-AP and AP groups from Experiment 2A.

Similar to Experiment 2A, we constructed a generalized linear model with a binomial link to assess performance differences between the AP and Expert groups when controlling for music experience – treated as a single, continuous variable with 5 levels. When controlling for music experience, performance in the AP group was significantly higher than performance in the Expert group ($\beta = 0.77$, $SE = 0.35$, $p = 0.03$). This group difference was not affected by adding self-reported familiarity with the censor tone into the model ($\beta = 0.77$, $SE = 0.35$, $p = 0.03$). Music experience was not a significant predictor of censor tone accuracy ($\beta = 0.11$, $SE = 0.13$, $p = 0.39$).

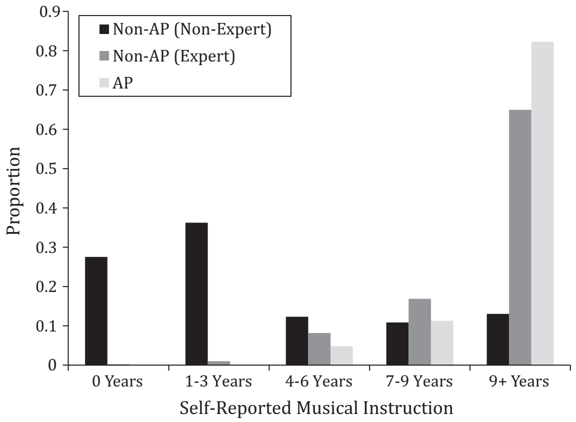

Fig. 3. Histogram of music experience across participant groups of Experiments 2A and 2B.

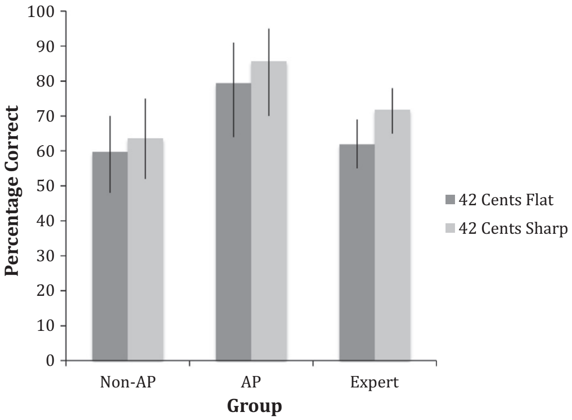

Fig. 4. Accuracy in selecting the correctly tuned censor tone among an alternative that was shifted 42 cents flat (dark grey) or 42 cents sharp (light grey) across the non-AP and AP possessor groups of Experiment 2A and the Expert group of Experiment 2B. All groups were, on average, above chance in selecting the correct censor tone (corresponding to 50%), though the AP group performed better than the non-AP group and Expert group, even when controlling for musical experience and stimulus familiarity. Error bars represent 95% credible intervals.

4.3. Discussion

Does long-term auditory memory for pitch, independent of explicit pitch labels, differ between AP possessors and non-AP possessors? If so, can this difference be explained by overall differences in musical experience? These questions were addressed by testing the memory of AP possessors, non-AP musical non-experts, and non-AP musical experts for the correct censor tone in situations where alternative versions of the tone were presented close enough in pitch (42 cents, or approximately a 2.5% difference) that note category information could not aid in the recognition decision.

The answers to these two primary questions are clear. First, there is evidence that non-AP possessors (both musical non-experts and experts) can still reliably choose the correctly tuned censor tone, even among alternatives that are very close in pitch (42 cents). To our knowledge, this is the first demonstration that long-term pitch memory is this fine-grained in a non-AP population. Second, the performance differences between AP and non-AP possessors in Experiment 1 are not likely entirely explained by overall differences in musical instruction, stimulus familiarity, or because AP possessors were using explicit note categories to make their judgments. When the incorrect version was flattened by 42 cents, making both the correct and incorrect versions within the same note category (21 cent flat "B" versus a 21 cent sharp "B"), AP possessors outperformed the non-AP sample by 19.7 points and outperformed the Expert sample by 17.5 points (with non-overlapping credible intervals).

The predicted improvement in performance among AP possessors when the incorrect version of the censor tone was sharpened by 42 cents was not statistically significant, though it was in the predicted direction (6.3% better performance for sharp conditions compared to flat conditions). This asymmetry is presumably caused because the sharp version of the censor tone traverses the putative note category boundary (21 cent sharp "B" versus a 37 cent flat "C"), whereas the flat version of the censor tone does not (21 cent sharp "B" versus a 21 cent flat "B"), which would have allowed the use of an explicit note category to differentiate the different versions (cf. Siegel, 1974). Interestingly, the Expert group also showed this asymmetry between flat and sharp conditions (9.9%, compared to 6.3% for AP possessors) which, given the larger sample size in the Expert group, was significant using a Fisher's Exact Test ($p = 0.04$, two-tailed). This asymmetry in the Expert group suggests either that the Experts had some limited knowledge of explicit note categories (cf. Miyazaki et al., 2012), or that regardless of *explicit* AP labeling ability, the Experts might have some implicit knowledge of note category boundaries. Despite this intriguing finding, the results from Experiments 2A and 2B suggest that AP possessors have enhanced pitch memories compared to non-AP possessors, independent of explicit note category knowledge and overall musical experience.

5. Experiment 3

The present experiment tested whether AP possessors' explicit category knowledge of the censor tone was precise enough to confer a performance advantage in Experiment 2. While it is possible, if not likely, that AP possessors used explicit category knowledge in Experiment 1, in which alternate tunings of the censor tone could differ by as much as two semitones, it seems much less likely that they were able to apply their prior explicit note category knowledge toward the censor tone in a manner that would be beneficial in Experiment 2. This is because the flat conditions involved differentiating censor tone tunings that mirrored an actual musical note – B5. Thus, in order for an AP possessors' explicit knowledge of note categories to provide an advantage in this case, they would need to have prior knowledge that the censor tone corresponds to a "sharp B." Indeed, if this were the case, then choosing the correct censor tone compared to an alternative that was a "flat B" would not necessarily rely on better pitch memory but on an AP possessor's ability to label notes.

5.1. Methods

5.1.1. Participants

We specifically recruited individuals with AP ($n = 61$) to participate in the experiment. Of these 61 participants, 42 reported being naïve to the general experimental paradigm (i.e., they reported no previous participation in an experiment where they were asked to judge the censor tone). The remaining 19 participants reported previously participating in a censor tone experiment. These participants completed different versions of the experiment, the details of which are outlined in the following sections.

5.1.2. Materials and procedure

The two sine tones that served as the different versions of the censor tone were generated in Adobe Audition (Adobe Systems: San Jose, CA). Unlike Experiments 1 and 2, there was no "correct" version (i.e., a sine tone generated at 1000 Hz). Rather, we generated two sine tones that mirrored the "correct" sine tone by 21 cents. Thus, given that 1000 Hz corresponds to an approximately 21-cent sharp "B," the lower tone we generated corresponded almost perfectly to an in-tune B5 (987.90 Hz), while the higher tone we generated corresponded almost perfectly to a 42-cent sharp B5 (1012.24 Hz).

The procedure differed based on whether a participant reported previously participating in a censor tone experiment. If participants reported no previous participation in a censor tone experiment (naïve AP participants), they were asked to think about the "bleep" that is used to censor taboo words on TV and on the radio. They were then asked to provide their best judgment of the note name *and* intonation of the censor tone. They could choose between all 12 note names, as well as the categories of "flat," "in-tune," and "sharp" for the intonation judgment. Importantly, these naïve participants did not hear any actual versions of the censor tone. We simply wanted to assess how accurately participants could explicitly label the censor tone using their absolute pitch categories.

Participants who reported previously participating in a censor tone experiment (non-naïve AP participants) were instructed that they would hear two versions of the "bleep" that is used to censor taboo words on TV and on the radio. However, it was emphasized that *neither* version corresponded perfectly to the censor tone they would encounter outside of the experiment. Thus, they were asked to choose the version they thought was most similar to the "correct" censor tone outside of the experiment. The ordering of the versions (in-tune "B" first vs. 42-cent sharp "B" first) was counterbalanced across participants. Apart from the tuning of the censor tones, the stimuli were identical to the ones used in the MT2 participants of Experiment 1 and all participants of Experiment 2 (using the synthetic talker). After making their judgment as to which version sounded most similar to the correct censor tone, participants were asked to label the 42-cent sharp "B" sine tone with its note name and intonation. They could choose between all 12 note names, as well as the categories of "flat," "in-tune," and "sharp" for the intonation judgment.

All participants were asked to provide an estimate of their musical instruction (in years). Music experience was defined as the number of years an individual had actively played a musical instrument, including the voice.

5.1.3. Participant inclusion

Of the 61 participants recruited in the AP group, everyone self-identified as possessing AP. Moreover, all participants were able to pass the short test of AP ability (which was identical to the one used in Experiments 1 and 2). We thus did not exclude any participant from our analyses.

5.2. Results

5.2.1. Naïve AP participants

If an explicit, preexisting note category can account for the AP performance advantage seen in Experiment 2, then AP possessors would need to have prior knowledge that the correct censor tone corresponds to a "sharp B." Both note category and intonation accuracy are important, since having a general sense of the note category (e.g., "some kind of B") would not allow an AP possessor to accurately distinguish between the correctly tuned censor tone (a "sharp B") and the flattened censor tone (a "flat B") used in Experiment 2.

The results from the naïve participants are displayed in Table 3. One thing that becomes immediately clear is that naïve participants, while generally accurate in applying a note category label to the censor tone, display a kind of perceptual magnet effect (cf. Athos et al., 2007) in which the censor tone – a 21-cent sharp "B" – is most frequently remembered as an "in-tune B." Importantly, given the way in which Experiment 2 was designed, a prior belief that the censor tone corresponds to an "in-tune B" would provide *no* performance advantage when selecting between the correct censor tone (21-cent sharp "B") and the flattened censor tone (21-cent flat "B"). While 4 out of 42 participants correctly labeled the censor tone as a "sharp B," which would confer a performance advantage in Experiment 2, more participants inaccurately believed the censor tone was a "flat B" (5 of 42) or a "flat C" (1 of 42), which would engender a performance *disadvantage* in Experiment 2, as these corresponded to the incorrect censor tone tunings. Thus, the results from the naïve AP participants suggest that while a slight majority of AP possessors (23 of 42, or 54.8%) are able to provide an accurate *note category* label for the censor tone, it does not appear that AP possessors contain the kind of intonation specificity within their explicit category that would explain the AP advantage found in Experiment 2.

5.2.2. Non-naïve AP participants

The results of the non-naïve AP participants similarly suggest that prior note and intonation category knowledge of the censor tone is not sufficient to explain the AP advantage found in Experiment 2. If participants heard the 42-cent "sharp B," compared to the "in-tune B," as most similar to the censor tone, then it suggests that perhaps they had used this distinction of "sharp B" when they had previously participated (i.e., either in Experiment 1 or 2). If, however, participants showed no preference of choice between the two censor tone versions, or if they consistently chose the "in-tune B" version as most similar to the correct censor tone, then it is unlikely they had used the category label of "sharp B" in a previous experiment.

Table 3
Intonation and note category estimates of the censor tone from naïve participants ($n = 42$). While AP possessors were generally accurate in remembering the censor tone note category (54.8% chose the correct note category and 85.7% were within one semitone), they did not appear to possess the level of specificity with regard to intonation that would explain the AP performance advantage in Experiment 2.

Number of participants	Intonation	Note category
0	Flat	A
2	In-Tune	A
1	Sharp	A
4	Flat	B-flat
3	In-Tune	B-flat
3	Sharp	B-flat
5	Flat	B
14	In-Tune	B
4	Sharp	B
1	Flat	C
2	In-Tune	C
0	Sharp	C
1	Flat	E-flat
1	In-Tune	E
1	Sharp	G-flat

The non-naïve AP participants showed a bias of choosing the in-tune "B" censor tone, which is similar to the expectations of the naïve AP listeners with respect to this tone. Of the 19 non-naïve AP participants, 14 (73.7%) chose the in-tune "B" as most closely corresponding to the correct censor tone, which, despite the relatively low sample size, was marginally significant (Sign Test: $p = 0.063$, two-tailed). However, this selection bias might be partly explained by how participants categorized the 42-cent sharp "B" with respect to a note name and intonation label. Specifically, if all participants heard the 42-cent sharp "B" as a "flat C," then they might have chosen the in-tune "B" simply because it was the only perceived "B" out of the two options. Thus, if virtually every participant heard the 42-cent sharp "B" as a "flat C," then it would remain unclear whether participants thought of the correct censor tone as a "sharp B."

Based on previous research, we would expect that a 42-cent sharp "B," being close to the putative 50-cent note category boundary, might be identified as a "sharp B" around half of the time, while being identified as a "flat C" around half of the time (e.g., see Fig. 1 from Levitin & Rogers, 2005). That being said, if a significantly greater number of participants labeled the 42-cent sharp "B" as a "flat C," it would interfere with our ability to conclude that non-naïve participants thought of the censor tone as more of an "in-tune B" versus a "sharp B." Of the 14 participants who selected the in-tune "B" as most similar to the correct censor tone, 5 classified the sharp "B" as a "sharp B," whereas 5 classified the sharp "B" as a "flat C." The remaining 4 classified the sharp "B" as a "flat B" ($n = 2$) and an "in-tune C" ($n = 2$). Of the 5 participants who selected the sharp "B" as most similar to the censor tone, 2 classified the sharp "B" as a "sharp B," while the remaining 2 classified the sharp "B" as a "flat B." It thus appears that the bias in selecting the in-tune "B" cannot be attributed to participants systematically mishearing the 42-cent sharp "B" as a "flat C," since there was close to an even split in how participants labeled the 42-cent sharp "B" (7 as "sharp B," 5 as "flat C").

5.3. Discussion

The purpose of Experiment 3 was to test the possibility that AP possessors might have outperformed non-AP possessors in Experiments 1 and 2 by using prior category knowledge of the censor tone as a "sharp B." We thus designed Experiment 3 to assess whether AP possessors used the label of "sharp B" with any consistency when encouraged to think about the censor tone in terms of its note category and intonation labels. If this pattern of results had been found it would support the conjecture that AP possessors outperformed non-AP possessors in the previous experiments through the use of explicit note and intonation categories rather than through enhanced pitch memory.

Across both naïve and non-naïve AP participants, we found converging evidence against this conjecture. Rather, evidence suggests that the censor tone was not consistently categorized as a "sharp B." If anything, we found evidence across both groups for a perceptual magnet effect in which the censor tone was most closely aligned with an "in-tune B" note category. Naïve AP participants were almost three times more likely to classify the censor tone as an "in-tune B" versus a "sharp B," and virtually equally likely to classify the censor tone as a "flat B" compared to a "sharp B." Non-naïve AP participants (from prior experiments) showed a bias toward thinking that an "in-tune B" was more similar to the correct censor tone compared to a 42-cent "sharp B," even though they showed no unexpected bias in explicitly labeling the 42-cent sharp "B" correctly (i.e. as a "sharp B"). Taken together, these results inform the AP advantage observed in Experiment 2A, particularly in the conditions where participants heard both the correct censor tone (21-cent "sharp B") and the flat censor tone (21-cent "flat B"), as conceptualizing the censor tone as an "in-tune B" would not confer any performance advantage in this situation.

6. General discussion

The present experiments demonstrate that pitch memory among non-AP possessors can be quite accurate, even for a "simple" stimulus that is non-musical and contains no harmonic information. Averaging across all analyzable participants from the MT1 and MT2 samples of Experiment 1 (total $n = 275$), we observe 73.6% accuracy in selecting the correct censor tone when the alternative version is shifted by one semitone and 83.6% accuracy when the alternative version is shifted by two semitones. Even when introducing alternative censor tones that were more fine-grained than a semitone (42 cents), non-AP possessors were reliably above chance (average of 65.6% accuracy) at selecting the correct version (total $n = 529$). Taken together, these results demonstrate that pitch memory among non-AP possessors can be remarkably accurate, even when using a stimulus that in many ways resembles a typical stimulus used to test for AP (i.e., an isolated pitch).

Regardless of whether the accuracy observed in the present experiments represents a qualitative difference between the censor tone and other operationalizations of pitch memory (e.g., music recordings), the important question is: How can pitch memory for an isolated tone with no harmonics be so accurate? There are at least three reasons. First, music recordings contain several salient dimensions to which an individual may attend, such as dynamic changes in the melody and timbre, thus interfering with the processing of absolute pitch information. Indeed, absolute and relative pitch have been described as competitive processes (e.g., Ramscar et al., 2011; Sergeant & Roche, 1973), which potentially suggests that the more relative pitch information a stimulus contains, the less likely the stimulus will be encoded absolutely. Second, the censor tone is highly salient and attention orienting. This is not to say that popular music recordings are not salient and attention orienting. Rather, the nature of the use of the censor tone – in which a speech signal is abruptly masked by a pure tone – likely results in "bottom-up" attentional capture (e.g., Kaya & Elhilali, 2014). In this sense, it is difficult *not* to attend to the censor tone when it is encountered in one's environment. Third, the censor tone is a highly stable environmental stimulus with respect to its absolute pitch. While popular music recordings are often heard with the same absolute pitches, there is almost certainly more variability when it comes to music, both in terms of how it is heard and produced (cf. Jakubowski & Müllensiefen, 2013). In contrast, the censor tone is highly unlikely to be produced by an individual, and is unlikely to be heard at a differing absolute pitch, at least in the United States.

Despite the relatively high performance across Experiments 1 and 2 for non-AP possessors, there was considerable evidence that AP possessors had even better pitch memories for the censor tone compared to the non-AP population. The notion that general pitch memory might be better in an AP versus a non-AP population has been previously claimed (Dooley, 2011), though it was unclear whether the reported performance difference was due to the ability of the AP population to use explicit note knowledge to perform the task. Even when controlling for several potential confounding factors in Experiment 2, including overall music experience, stimulus familiarity, and the use of explicit note category knowledge as an effective source of information, we found evidence that AP possessors were still outperforming non-AP possessors. Further, Experiment 3 provided consistent evidence that AP possessors did not have the necessary note and intonation category specificity to effectively use an explicit category label in Experiment 2.

How can this AP advantage be interpreted in the larger framework of absolute pitch memory formation? One intriguing possibility is that individuals who lie on the high end of the pitch memory distribution (cf. Schellenberg & Trehub, 2003) might be more prone to developing genuine AP, given the right kind of musical instruction. In other words, individuals with AP might be generally better at forming precise long-term auditory memories, which then leads to the ability to form explicit long-term AP categories given the right kind of environmental input. This interpretation of the present results largely fits within the framework of David Ross and colleagues, who have argued that AP possessors have a fundamentally different way of absolutely perceptually encoding (APE) incoming sounds, which is dissociable from the existence and use of music note labels (e.g., see Ross et al., 2005). However, while Ross et al. (2005) interpret this category-independent AP advantage as evidence for the "innate" nature of APE, we would argue for a different explanation. Specifically, given the continuous nature of general pitch memory among non-AP possessors (cf. Schellenberg & Trehub, 2003) and AP ability among true AP possessors (e.g., Bermudez & Zatorre, 2009), it is possible that individual differences in "APE" reflect a more continuous difference in pitch memory, which can be applied with varying success in the precision of explicit long-term categories (whether those categories happen to be musical note names or something else, like the "censor tone").

Along these lines, recent work has demonstrated that among non-AP possessors, the ability to hold onto pitch information in working memory predicts the success of explicitly learning long-term AP categories (Van Hedger, Heald, Koch, & Nusbaum, 2015). Another potential reason to support the notion that a natural enhancement in forming auditory memory representations leads to AP comes from Deutsch and Dooley (2013), who found that AP possessors appear to have better auditory digit spans compared to musically matched controls. Given that this memory advantage is shown outside of the realm of music (and it is unlikely that explicit note category knowledge would help differentiate subtle pitch differences between spoken numbers), this suggests that perhaps AP develops in part as a function of an enhanced general ability to remember auditory information (though it is currently unclear how this kind of memory advantage relates to "absolute perceptual encoding" or working memory for pitch).

Another possible explanation for the AP advantage observed in the present experiments is that AP possessors may have more precise auditory long-term memories for pitch as a function of their explicit long-term note categories. Specifically, pitch memory, if fine-grained enough, would allow for plasticity in an explicit AP labeling system (cf. Gureckis & Goldstone, 2008). Put more simply, a robust memory for fine-grained pitch differences would sensitize AP possessors to environmental shifts in note tunings, affording them the ability to adapt their labeling behavior. Given that our recent work has shown that experience with slightly detuned music can shift the tunings of an AP possessor's explicit labeling system (Hedger et al., 2013), it is possible that the pitch memory performance enhancements in the AP population compared to the non-AP populations is due to the role that long-term pitch memory precision may play instantiating such flexibility.

This is consistent with the model of categorization proposed by Huttenlocher, Hedges, and Vevea (2000) for other domains (e.g., length and size estimation). In particular, Huttenlocher et al. (2000) argue that category level information influences individuals' stimulus reproductions in a systematic manner that maximizes accuracy. In audition, recent work has shown that while all participants – regardless of musical experience – held some generalized note category knowledge that helped them accurately reproduce auditory pitches, AP possessors showed advantages in reproducing pitches corresponding to a musical note (C5) that could not be attributed to music experience (Heald, Van Hedger, & Nusbaum, 2014). To be clear, this theoretical explanation of the AP advantage observed in the present set of experiments is different than the idea that AP possessors are performing better simply because they are *using* their explicit long-term pitch category knowledge to perform the task (as this possibility was made unlikely through the design of Experiments 2 and 3). Rather, the implication of this alternative interpretation is that AP possessors – through their explicit category expertise – are able to generally represent auditory pitch in a superior manner compared to non-AP possessors, even in instances where stimuli cannot be differentiated at the note category level and consequently the level of specificity afforded by explicit categories does not provide a performance advantage. While future work is necessary to make a stronger causal claim between the relationship of general pitch memory and explicit AP ability, the present results suggest that independent of explicit AP knowledge, overall music experience, and stimulus familiarity, AP possessors have better long-term pitch memory compared to non-AP possessors, at least for long-term pitch categories.

Performance in the present study is particularly notable considering the fact that the censor tone is a sine tone, which has consistently proven to be among the most difficult timbres for genuine AP possessors to accurately identify (e.g., Lee, Lee, & Shr, 2011; Lockhead & Byrd, 1981; Miyazaki, 1989; though see Vanzella & Schellenberg, 2010). There are two main reasons why sine tones have been thought to be particularly difficult to identify. First, sine tones by definition do not have harmonic structures, which can aid in pitch perception. This explanation, however, does not seem to be sufficient in explaining "sine tone deficits" among AP possessors, especially since sine tones in the range of 1000 Hz are actually just as discriminable, if not more discriminable, compared to complex tones (e.g., Oxenham & Micheyl, 2013). The second possible explanation of the "sine tone deficit" is that sine tones are relatively uncommon in the environment. Given that AP ability is at least partly experience dependent, with individuals performing faster and/or more accurately for frequently encountered instrumental timbres, pitch ranges, and individual notes (e.g., Bahr, Christensen, & Bahr, 2005; Miyazaki, 1989), the relative lack of familiarity with sine tones is a particularly appealing explanation for the "sine tone deficits" often observed in tests of AP ability. While the present set of experiments cannot directly address this peculiarity in (mis)classifying sine tones among AP possessors, as the experimental paradigm was simplified relative to traditional tests of AP, the results from both experiments suggest that sine tone accuracy can be remarkably high when providing adequate context (i.e. boosting familiarity).

In conclusion, despite the impressive pitch memory for the censor tone among non-AP possessors, we found consistent evidence that AP possessors were overall more accurate at selecting the correct censor tone. This advantage could not be explained through overall differences in musical instruction or through the possibility that AP possessors had an explicit representation of the censor tone that contained the necessary note and intonation category information to confer a performance advantage. Thus, these results support the notion that true AP ability is differentiated at a level other than explicit note category knowledge. While more research is needed to expand upon this notion, these results point to the possibility that general differences in long-term auditory memory might be responsible for the development of AP, or that the development of long-term explicit AP categories affects the subsequent fidelity of auditory memory.

Acknowledgements

This research was supported in part by a grant from the ONR grant DoD/ONR N00014-12-1-0850 to UCSD and a departmental fellowship (Norman H. Anderson Fund). The data from these experiments can be accessed on the Open Science Framework (https://osf.io/4e6uy).

References

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 of the United States of America, 104*(37), 14795–14800. http://dx.doi.org/10.1073/pnas.0703868104.

Bååth, R. (2014). Bayesian first aid: A package that implements bayesian alternatives to the classical ⁎.test functions in R. In *User! 2014. Los Angeles, CA.* .

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

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

Bahr, N., Christensen, C. A., & Bahr, M. (2005). Diversity of accuracy profiles for absolute pitch recognition. *Psychology of Music, 33*(1), 58–93. http://dx.doi.org/10.1177/0305735605048014.

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

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. http://dx.doi.org/10.1121/1.3081389.

Deutsch, D., & Dooley, K. (2013). Absolute pitch is associated with a large auditory digit span: A clue to its genesis. *The Journal of the Acoustical Society of America, 133*(4), 1859–1861.

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), 719–722. http://dx.doi.org/10.1121/1.2151799.

Dooley, K. (2011). *Absolute pitch and related abilities*. permalink: UC San Diego. http://escholarship.org/uc/item/7j46q4z5.

Frieler, K., Fischinger, T., Schlemmer, K., Lothwesen, K., Jakubowski, K., & Mullensiefen, D. (2013). Absolute memory for pitch: A comparative replication of Levitin's 1994 study in six European labs. *Musicae Scientiae, 17*(3), 334–349. http://dx.doi.org/10.1177/1029864913490633.

Fujisaki, W., & Kashino, M. (2002). The basic hearing abilities of absolute pitch possessors. *Acoustical Science and Technology, 23*(2), 77–83. http://dx.doi.org/10.1250/ast.23.77.

Gureckis, T., & Goldstone, R. L. (2008). The effect of the internal structure of categories on perception. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 843). Austin, TX: Cognitive Science Society.

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

Heald, S. L. M., Van Hedger, S. C., & Nusbaum, H. C. (2014). Auditory category knowledge in experts and novices. *Frontiers in Neuroscience, 8*(August), 1–15. http://dx.doi.org/10.3389/fnins.2014.00260.

Hedger, S. C., Heald, S. L. M., & Nusbaum, H. C. (2013). Absolute pitch may not be so absolute. *Psychological Science, 24*(8), 1496–1502.

Hutchins, S., Hutka, S., & Moreno, S. (2015). Symbolic and motor contributions to vocal imitation in absolute pitch. *Music Perception: An Interdisciplinary Journal, 32*(3), 254–265. http://dx.doi.org/10.1525/mp.2015.32.3.254.

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General, 129*(2), 220–241. http://dx.doi.org/10.1037/0096-3445.129.2.220.

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 (2006), 66*(7), 1259–1267. http://dx.doi.org/10.1080/17470218.2013.803136.

Kaya, E. M., & Elhilali, M. (2014). Investigating bottom-up auditory attention. *Frontiers in Human Neuroscience, 8*(May), 327. http://dx.doi.org/10.3389/fnhum.2014.00327.

Kishon-Rabin, L., Amir, O., Vexler, Y., & Zaltz, Y. (2001). Pitch discrimination: Are professional musicians better than non-musicians? *Journal of Basic and Clinical*

Physiology and Pharmacology, 12(2), 125–143. http://dx.doi.org/10.1515/JBCPP.2001.12.2.125.

Lee, C. Y., Lee, Y. F., & C.L., Shr (2011). Perception of musical and lexical tones by Taiwanese-speaking musicians. *The Journal of the Acoustical Society of America, 130*(1), 526–535. http://dx.doi.org/10.1121/1.3596473.

Levitin, D. J. (1994). Absolute memory for musical pitch: Evidence from the production of learned melodies. *Perception & Psychophysics, 56*(4), 414–423. http://dx.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. http://dx.doi.org/10.1016/j.tics.2004.11.007.

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

Micheyl, C., Delhommeau, K., Perrot, X., & Oxenham, A. J. (2006). Influence of musical and psychoacoustical training on pitch discrimination. *Hearing Research, 219*(1–2), 36–47. http://dx.doi.org/10.1016/j.heares.2006.05.004.

Miyazaki, K. (1989). Absolute pitch identification: Effects of timbre and pitch region. *Music Perception, 7*(1), 1–14. http://dx.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. http://dx.doi.org/10.1121/1.4756956.

Oxenham, A. J, & Micheyl, C. (2013). Pitch perception: Dissociating frequency from fundamental-frequency discrimination. In B. C. J. Moore et al. (Eds.). *Basic aspects of hearing* (Vol. 787, pp. 137–145). New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4614-1590-9_16.

Profita, J., & Bidder, G. (1988). Perfect pitch. *American Journal of Medical Genetics, 29*, 763–771. http://dx.doi.org/10.1016/S1364-6613(00)01642-9.

Rakowski, A., & Morawska-Bungeler, M. (1987). In search for the criteria of absolute pitch. *Archives of Acoustics, 12*, 75–87.

Rakowski, A. (n.d.). Direct comparison of absolute and relative pitch. In F.A. Bilsen (Ed.), *Symposium on hearing theory* (pp. 105–108). Eindhoven, The Netherlands: Instituut voor Perceptie Underzoek.

Ramscar, M., Suh, E., Dye, M., 2011. How pitch category learning comes at a cost to absolute frequency representations. In 33rd Meeting of the Cognitive Science Society. Boston, MA.

Ross, D. A., Gore, J. C., & Marks, L. E. (2005). Absolute pitch: Music and beyond. *Epilepsy and Behavior, 7*(4), 578–601. http://dx.doi.org/10.1016/j.yebeh.2005.05.019.

Ross, D. A., & Marks, L. E. (2009). Absolute pitch in children prior to the beginning of musical training. *Annals of the New York Academy of Sciences, 1169*, 199–204. http://dx.doi.org/10.1111/j.1749-6632.2009.04847.x.

Ross, D. A., Olson, I. R., & Gore, J. C. (2003). Absolute pitch does not depend on early musical training. *Annals of the New York Academy of Sciences, 999*, 522–526. http://dx.doi.org/10.1196/annals.1284.065.

Schellenberg, E. G., & Habashi, P. (2015). Remembering the melody and timbre, forgetting the key and tempo. *Memory & Cognition, 43*(7), 1021–1031. http://dx.doi.org/10.3758/s13421-015-0519-1.

Schellenberg, E. G., Stalinski, S. M., & Marks, B. M. (2014). Memory for surface features of unfamiliar melodies: Independent effects of changes in pitch and tempo. *Psychological Research Psychologische Forschung, 78*(1), 84–95. http://dx.doi.org/10.1007/s00426-013-0483-y.

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

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

Sergeant, D., & Roche, S. (1973). Perceptual shifts in the auditory information processing of young children. *Psychology of Music, 1*(2), 39–48.

Siegel, J. A. (1974). Sensory and verbal coding strategies in subjects with absolute pitch. *Journal of Experimental Psychology, 103*, 37–44. http://dx.doi.org/10.1121/1.1919544.

Smith, N. A., & Schmuckler, M. a. (2008). Dial A440 for absolute pitch: Absolute pitch memory by non-absolute pitch possessors. *The Journal of the Acoustical Society of America, 123*(4), EL77–EL84. http://dx.doi.org/10.1121/1.2896106.

Takeuchi, A. H., & Hulse, S. H. (1993). Absolute pitch. *Psychological Bulletin, 113*(2), 345–361. http://dx.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: An Interdisciplinary Journal, 1*(1), 63–83. http://dx.doi.org/10.2307/40285250.

Van Hedger, S. C., Heald, S. L. M., Koch, R., & Nusbaum, H. C. (2015). Auditory working memory predicts individual differences in absolute pitch learning. *Cognition, 140*, 95–110.

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

Ward, W. D. (1999). Absolute pitch. In D. Deutsch (Ed.), *The psychology of music* (2nd ed., pp. 265–298). San Diego, CA: Academic Press.