Cognition 218 (2022) 104949

Contents lists available at ScienceDirect

Cognition

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

Musical instrument familiarity affects statistical learning of tone sequences

Stephen C. Van Hedger [a,b,c,*], Ingrid S. Johnsrude [a,b,d,e], Laura J. Batterink [a,b]

[a] Department of Psychology, University of Western Ontario, London, ON, Canada
[b] Brain and Mind Institute, University of Western Ontario, London, ON, Canada
[c] Department of Psychology, Huron University College, London, ON, Canada
[d] National Centre for Audiology, University of Western Ontario, London, ON, Canada
[e] School of Communication Sciences and Disorders, University of Western Ontario, London, ON, Canada

ARTICLE INFO

Keywords:
Statistical learning
Generalization
Music
Expertise
Familiarity
Perception

ABSTRACT

Most listeners have an implicit understanding of the rules that govern how music unfolds over time. This knowledge is acquired in part through statistical learning, a robust learning mechanism that allows individuals to extract regularities from the environment. However, it is presently unclear how this prior musical knowledge might facilitate or interfere with the learning of novel tone sequences that do not conform to familiar musical rules. In the present experiment, participants listened to novel, statistically structured tone sequences composed of pitch intervals not typically found in Western music. Between participants, the tone sequences either had the timbre of artificial, computerized instruments or familiar instruments (piano or violin). Knowledge of the statistical regularities was measured as by a two-alternative forced choice recognition task, requiring discrimination between novel sequences that followed versus violated the statistical structure, assessed at three time points (immediately post-training, as well as one day and one week post-training). Compared to artificial instruments, training on familiar instruments resulted in reduced accuracy. Moreover, sequences from familiar instruments – but not artificial instruments – were more likely to be judged as grammatical when they contained intervals that approximated those commonly used in Western music, even though this cue was non-informative. Overall, these results demonstrate that instrument familiarity can interfere with the learning of novel statistical regularities, presumably through biasing memory representations to be aligned with Western musical structures. These results demonstrate that real-world experience influences statistical learning in a non-linguistic domain, supporting the view that statistical learning involves the continuous updating of existing representations, rather than the establishment of entirely novel ones.

1. Introduction

Imagine that you have been tasked with visiting an alien civilization to learn its language, customs, and cultural practices. Upon arriving, you are bombarded with completely unfamiliar sights and sounds, with no translator to help you make sense of your new environment. How can you hope to learn anything in such a seemingly impossible situation? This farfetched scenario resembles the challenges faced by infants, who must figure out a new system of complex rules and regularities without a priori knowledge of what environmental features are informative. Statistical learning – commonly defined as the process of becoming sensitive to patterns in the environment – has been proposed as a powerful and domain-general learning mechanism that allows individuals to implicitly pick up on these kinds of regularities in their environments (e.g., see Saffran & Kirkham, 2018). This sensitivity to statistical regularities allows learners to discover higher-order structure in an unsupervised manner (Fiser & Aslin, 2001), and has been proposed to underlie the learning of regularities in both language and music (McMullin & Saffran, 2004). The seminal observation that eight-month-old infants can track statistical patterns in a stream of continuous speech (Saffran, Aslin, & Newport, 1996) and musical tones (Saffran, Johnson, Aslin, & Newport, 1999) further highlights the potential importance of statistical learning in the initial development of relevant units in both language (e.g., words and syntax) and music (e.g., melodic and harmonic sequences).

Given that statistical learning facilitates the learning of perceptual structures in the environment, researchers studying statistical learning often design their stimuli to be *novel* (i.e., unlikely to have been

* Corresponding author at: Department of Psychology, Huron University College, London, ON, Canada.
E-mail address: svanhedg@uwo.ca (S.C. Van Hedger).

https://doi.org/10.1016/j.cognition.2021.104949
Received 25 June 2020; Received in revised form 22 June 2021; Accepted 28 October 2021
Available online 9 November 2021

encountered outside of the experimental context), to minimize the possibility of prior learning influencing the results. For example, linguistic statistical learning paradigms typically involve presenting learners with sequences composed of trisyllabic nonsense words that do not resemble existing words in the learners' own language (e.g., *bidaku*). Learning is then assessed by a post-exposure test, in which participants are asked to discriminate between target items (from the exposure period) and foil items, composed of individual units presented in a different order (e.g., Saffran et al., 1996). In non-linguistic statistical learning paradigms that use tone sequences, the relative pitches in the sequence are often carefully selected to not overlap with any Western musical scales, in order to reduce the possibility that previously acquired knowledge will systematically influence learning (e.g., Loui, 2012; Loui & Wessel, 2008; though see Saffran et al., 1999). Similarly, visual statistical learning paradigms often use abstract geometric shapes that cannot be readily labeled in order to minimize the role of prior experience (Turk-Browne, Junge, & Scholl, 2005).

However, despite these careful considerations to make the items *novel* in statistical learning paradigms, prior knowledge may still have an impact on learning outcomes. Listeners in statistical learning experiments unavoidably have extensive expertise in relevant domains, such as language and music (e.g., Bigand & Poulin-Charronnat, 2006; Ettlinger, Margulis, & Wong, 2011; Rohrmeier & Rebuschat, 2012). Even infant listeners cannot be viewed as entirely naïve to the underlying statistical structures of their environments (e.g., Kuhl, 2000, 2004; Werker & Tees, 1984, 1999). Although researchers' use of novel sequences in statistical learning tasks is an attempt to minimize the role of pre-existing memory representations, the stimuli to be learned may still overlap considerably with individuals' prior experiences accumulated outside of the lab.

In particular, linguistic statistical learning paradigms have displayed some puzzling patterns of results that suggest prior knowledge does indeed influence learning (Frost, Armstrong, Christiansen, & Armstrong, 2019). For example, Siegelman and Frost (2015) found that individuals displayed stable statistical learning performance within separate visual and auditory tasks, yet performance *across* modalities and tasks was essentially uncorrelated. Such independence of learning across domains is especially perplexing given that statistical learning is commonly described as a domain-general learning mechanism (e.g., Saffran & Kirkham, 2018), though the extent to which statistical learning is considered a unified learning mechanism is debated (e.g., see Endress & Bonatti, 2016; Thiessen, 2017; Zhao, Ngo, McKendrick, & Turk-Browne, 2011). Perhaps even more striking, Erickson, Kaschak, Thiessen, and Berry (2016) examined auditory statistical learning using different "languages" (syllable collections) and found stable test-retest performance within a language, yet the correlation between languages was surprisingly low and not significant. Finn and Hudson Kam (2008) also showed that statistical learning did not occur if the nonsense words in the speech stream were phototactically impossible in English. Once again, such large syllable-based effects suggest that learning is more context dependent than typically thought. Prior linguistic knowledge has also been shown to interfere with other aspects of language learning. For instance, Finn & Hudson Kam (2015) found that adults' native language knowledge interferes with learning morphological variation in a novel, unsegmented language. In this sense, prior knowledge of phonotactics can have a cascading impact, influencing word segmentation and the perception of morphological variation that relies upon proper segmentation.

Siegelman, Bogaerts, Elazar, Arciuli, and Frost (2018) propose that "entrenchment" may account for these results. The principle of entrenchment suggests that the validity of *tabula rasa* assumptions in a statistical learning paradigm may depend on how closely the learning task engages prior knowledge. In the case of learning sequences of abstract geometric shapes or novel fractal images (Schapiro, Turk-Browne, Norman, & Botvinick, 2016; Turk-Browne et al., 2005), the *tabula rasa* assumptions may be more valid, as individuals do not typically encounter sequences of discrete, static geometric patterns outside of experimental contexts. In contrast, in the case of learning novel linguistic structures (e.g., Saffran et al., 1996), adult listeners have already accrued vast amounts of linguistic experience outside of the experiment and thus may show more variable or idiosyncratic performance compared to a less "entrenched" domain. For example, some novel "words" in an artificial language may closely align with the phonotactic regularities of the participant's native language, and thus be more easily learned than items that do not conform to the native language. Prior knowledge and expectations about linguistic units thus appear to have a measurable influence on statistical learning performance.

Despite growing evidence that prior knowledge influences language-based statistical learning tasks, it is presently unclear whether statistical learning in *non-linguistic* domains is similarly influenced by prior knowledge. This question is of theoretical and practical importance to the understanding of statistical learning. If statistical learning is influenced by prior knowledge in a non-linguistic domain, this suggests that statistical learning may be more broadly conceptualized as a process of continuously updating acquired representations. In other words, the assumption that statistical learning reflects "isolated" learning, as opposed to some interaction between learning and pre-existing long-term memory representations, may only be tenable under limited and artificial experimental circumstances. Furthermore, examining how prior knowledge influences statistical learning of non-linguistic stimuli will also inform the degree to which findings from language paradigms generalize to other domains.

In the present study, we tested whether prior knowledge influences statistical learning performance outside of a linguistic context by using sequences of discrete tones that may invoke listeners' prior musical knowledge. Music is an excellent parallel system to language to assess the domain generality of "entrenchment" in statistical learning. Both theoretical and empirical work has demonstrated that listeners are sensitive to the statistical input of their musical environments, just as in the speech domain (McMullin & Saffran, 2004). In fact, the observation that listeners show similar performance in statistical learning and artificial grammar learning paradigms when linguistic stimuli are replaced with tones was foundational to the understanding of statistical learning as a more domain-general process (e.g., Loui, Wessel, & Hudson Kam, 2010; Saffran et al., 1999). Additionally, listeners typically develop a rich understanding of musical structures *implicitly* – i.e., from mere exposure, without explicit musical training (Tillmann & Bigand, 2000). For example, listeners can implicitly learn several musical features, including timbre and melodic pitch relationships, from their listening environments (e.g., see Rohrmeier & Rebuschat, 2012).

Another important parallel between language and music is that perceptual and memory processes within both systems are strongly shaped by experience. In speech, infants display an initial ability to learn phonetic contrasts found across all languages, but this perceptual sensitivity gradually becomes tuned specifically to the phonetic contrasts experienced within the first year of development (Kuhl, Williams, Lacerda, Stevens, & Lindblom, 1992; Stager & Werker, 1997; Werker & Tees, 1984). This perceptual narrowing can result in an inability to perceive specific contrasts that are not used in one's primary language (e.g., differentiating/r/and/l/in Japanese). Once adulthood has been reached, even extensive perceptual training generally produces only modest improvements in perceptual classification (Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Lim & Holt, 2011; Lively, Logan, & Pisoni, 1993). Similar processes have been described in music, specifically in the domains of pitch and rhythm perception. Western infants are equally adept at detecting whether a musical sequence contains an incorrect (mistuned) note, regardless of whether they are listening to Western or Javanese scales; in contrast, Western adult listeners are significantly better at detecting deviant notes in the context of Western scales (Lynch et al., 1990; Lynch & Eilers, 1992). Infants additionally display early preferences for the musical meter of their own culture (Soley & Hannon, 2010) and patterns of culture-specific learning that

emerges over the first year of life (Hannon, Soley, & Ullal, 2012; Hannon & Trehub, 2005a, 2005b). Importantly, similar to perceptual narrowing in speech perception, the culture-specific knowledge of adults appears to be extensive and might interfere with the learning of novel rhythmic information (Hannon & Trehub, 2005b). Taken together, these results suggest that our perceptual experiences of the world are continually narrowed and refined to align with the experiences we have accumulated in both speech and music. Thus, the extent to which statistical learning sequences overlap with this prior knowledge should influence performance in both language- and tone-based statistical learning paradigms.

However, experimental variability across previous studies has made it difficult to systematically assess how prior knowledge and experience influence tonal sequence learning, if at all. Specifically, the constituent tone frequencies sometimes adhere to conventional musical note values (e.g., Saffran et al., 1999) and sometimes are specifically tuned to avoid typical musical intervals (e.g., Loui et al., 2010). Most previous studies use timbres that are generally unfamiliar in musical contexts, such as sine tones (Loui & Wessel, 2008; Saffran et al., 1999), yet some studies use familiar timbres, such as that of a piano (e.g., Kuhn & Dienes, 2005; Leung & Dean, 2018). These inter-study differences make it difficult to quantify the relative effects of prior knowledge on learning efficacy.

To address this issue, the present experiment directly assessed whether statistical learning of tone sequences is influenced by prior musical experience. If so, we would expect that the *extent of overlap* or resemblance between experimental stimuli and musical elements heard outside the laboratory should influence learning. Specifically, we manipulated the timbre of the individual tones to either be unfamiliar (computer-synthesized specifically for this experiment, and unlikely to have been previously encountered) or highly familiar in a musical context (piano or violin).

1.1. Importance of timbre in music processing

Timbre is defined as the quality of a musical note that is not loudness or pitch; it is the quality that, for example, allows one to distinguish a middle "C" played on a piano from one played on an oboe. Several converging results suggest that timbre is an integral part of how tonal information is perceived (e.g., Hutchins, Roquet, & Peretz, 2012) and remembered in melodic contexts (Schellenberg & Habashi, 2015; Weiss, Trehub, & Schellenberg, 2012), suggesting that timbre might modulate the extent to which prior musical representations are brought online. First, expert musicians show enhanced perception for their primary instruments, which is accompanied by both cortical and subcortical neurophysiological changes (Margulis, Mlsna, Uppunda, Parrish, & Wong, 2009; Strait, Chan, Ashley, & Kraus, 2012) and is driven in part by auditory-motor interactions (Lahav, Boulanger, Schlaug, & Saltzman, 2005). Second, individuals with absolute pitch – the rare ability to name or produce a musical note without a reference (Takeuchi & Hulse, 1993) – consistently show timbre-specific effects in terms of speed and accuracy of note identification, with familiar timbres (e.g., piano) identified significantly faster and more accurately than other complex timbres (Bahr, Christensen, & Bahr, 2005; Miyazaki, 1989; Van Hedger & Nusbaum, 2018; Vanzella & Schellenberg, 2010). Notably, even individuals without extensive musical training or absolute pitch have demonstrated some influence of instrumental timbre on judging whether an instrument is "in tune." Musicians and non-musicians alike show a "vocal generosity effect," characterized by more lenient judgments of tuning for notes with the timbre of a singing voice compared to a violin (Hutchins et al., 2012). Most listeners are also able to judge when an isolated note is "in tune" according to conventional Western tuning standards, *though only when judging familiar timbres*; complex but non-musical timbres were not reliably differentiated based on tuning (Van Hedger, Heald, Huang, Rutstein and Nusbaum, 2017). Taken together, these findings indicate that instrumental timbre invokes specific prior knowledge, influencing perceptual and memory processes, and thus may also influence statistical learning of novel tone sequences.

1.2. Overview of design and predictions

In the present experiment, we adopted the paradigm outlined by Durrant, Taylor, Cairney, and Lewis (2011), which uses tonal intervals that are specifically designed to not overlap with previously learned musical categories. However, unlike Durrant and colleagues, we manipulated the timbre of the sequences. In our experiment, timbre was manipulated between participants such that individual tones had completely novel timbres (synthesized specifically for this experiment) or highly familiar timbres (produced by a standard musical instrument, either piano or violin). By keeping all other aspects of the experiment identical, we directly tested whether familiarity of a surface-level stimulus feature – in this case, timbre - influenced statistical learning performance.

All participants were exposed to a ~ 7 min structured sequence of tones during training. After training, learning was tested via a two-alternative forced-choice task, in which participants heard two shorter tonal sequences on each trial and determined which one sounded grammatical (i.e., more similar to the longer sequence heard in training). Participants additionally rated their confidence in their answers. To determine the robustness of learning, we manipulated the extent to which the test sequences resembled those heard in training, in two primary ways. The first way was in terms of perceptual tone similarity. Specifically, test sequences could be composed of individual tones that were acoustically identical to those in the training sequence ("Specific Test") or that varied in pitch and/or timbre ("Transfer Test"). The second way was in terms of grammatical adherence to the trained probabilistic structure, resulting in two levels of "Difficulty." Specifically, some sequences contained a high number of expected tone transitions (comparable to the training sequence) whereas other sequences contained a relatively fewer number of expected tone transitions and thus can be considered more challenging assessments of learning, allowing us to assess participant sensitivity to the learned structure under "noisier" conditions. These learning tests were repeated three times: (1) immediately after training, (2) one day after training, and (3) at least one week after training.

We hypothesized that instrument familiarity should influence statistical learning performance, given the integral relationship between instrumental timbre and the activation of specific musical representations, although the effect could go in either direction. It is possible that familiar instruments might *improve* overall statistical learning performance, similar to findings outside the music domain showing that source familiarity improves performance (e.g., familiar talkers leading to better intelligibility for novel linguistic phrases; Holmes, Domingo, & Johnsrude, 2018; Johnsrude et al., 2013; Nygaard & Pisoni, 1998). In the context of the present experiment, it could be the case that familiar instruments allow listeners to scaffold upon existing representations to better encode and remember the novel sequences, despite the fact that the sequences are designed to not overlap with typical musical intervals. In the present paradigm, this could manifest in terms of higher overall accuracy for participants trained with familiar instruments, in addition to better generalization (e.g., to a different frequency range) and better retention of learning over time.

Alternatively, it is possible that familiar instruments may *attenuate* overall statistical learning performance. This could occur because the use of familiar instruments might more strongly evoke previously learned musical structures, ultimately resulting in the *misperception* of the novel tonal sequences. This misperception could manifest in the moment of hearing the sounds, which would be consistent with a framework of categorical perception (e.g., Liberman & Mattingly, 1985; Siegel & Siegel, 1977). However, even in the absence of classic categorical perception effects, listeners might form weaker and/or biased long-term memory representations of the true intervallic relationships when played on a familiar instrument, as the memory representations

would be more likely to be warped to align with Western musical structures in the time since training (cf. Bartlett, 1932). In contrast, the artificial instruments – which were generated specifically for the purpose of this experiment – might not activate categorical representations for listeners, as these sounds presumably would not evoke prior musical representations to the same extent as familiar instruments. In the present paradigm, this could manifest in terms of lower overall accuracy for participants trained with familiar instruments, in addition to worse generalization and a more severe loss of learning in the time since training. Additionally, the association between confidence ratings and accuracy might be weakened, ostensibly because participants might conflate instrument familiarity with sequence familiarity. Finally, if familiar instruments lead to misperceiving the novel tonal sequences to be in line with existing musical representations, participants may be more likely to judge a sequence as grammatical if it happens to contain elements that resemble familiar music (such as familiar Western intervals), even if these cues are orthogonal to the grammatical structure.

If statistical learning performance is influenced by instrument familiarity, this would advance our understanding of how prior knowledge influences statistical learning in two ways. First, it would suggest that knowledge entrenchment in statistical learning is not limited to linguistic paradigms. Second, it would suggest that prior knowledge is activated dynamically based on the extent to which the statistical learning stimuli overlap with previously acquired knowledge structures (i.e., implying that the influence of prior knowledge can be modulated by context – in this case, the timbre of the sequence). In other words, it would suggest that timbral cues are a critical component for learning, not simply a surface-level feature that is stripped away in service of forming a more abstract memory representation.

2. Method

2.1. Participants

We recruited 200 participants through Amazon's Mechanical Turk (MTurk), an online crowdsourcing platform. Participants were recruited via Turk Prime, which was recently rebranded as CloudResearch (Litman, Robinson, & Abberbock, 2017). CloudResearch is a service that interfaces with MTurk and allows for additional participant recruitment control. Participants had to be residing in the United States or Canada and had to have a minimum 90% approval rating from at least 100 prior assignments to qualify. All participants gave consent through checking a box on the computer screen and were compensated for their participation. Participants were assigned either to the Artificial Instrument Training Group ($n = 100$), which used an unfamiliar timbre (see below), or the Familiar Instrument Training Group ($n = 100$), determined by the order run. Both conditions were run within six weeks of each other, with the Artificial Instrument Training Group run first (early June 2019) and the Familiar Instrument Training Group run second (mid July 2019). Two participants' data in the Familiar Instrument condition were not successfully sent to the data server, resulting in 98 participants. The research protocol was approved by Western University's Non-Medical Research Ethics Board.

2.2. Tone sequence description

The sequences used in the present experiment were modeled on those described in Durrant et al. (2011) and consisted of intervals not heard in Western tonal music. In contrast to Western music, in which an octave is divided into *twelve* equal steps, our experimental tonal system divided the octave into *five* equal steps. The difference between individual notes can be quantified by a logarithmic unit known as *cents*. Adjacent notes in Western music are separated by 100 cents, whereas adjacent notes in our experimental tonal system are separated by 240 cents. Situating this interval size in the context of Western music, if one were to establish the starting note of this novel tuning system as an "A," the next note in the system would be unplayable (at least on an instrument like the piano), as it would fall between the notes B and C.

These five tones formed the foundation of the statistical sequences. Similar to Durrant et al. (2011), the sequences were structured with a second-order Markov chain. This means that the prior *two* tones of the sequence determined the probability of hearing the next tone. The longer tone sequences used in training adhered to a high probability transition matrix (Fig. 1A), in which the prior two tones were highly predictive of the third tone (90%). For example, if the tone sequence started with repeats of tone 1 (i.e., the first row in Fig. 1A), then tone 4 would be 90% likely to occur. If tone 4 indeed occurred, then there would be a 90% probability of hearing tone 1 next (as the prior two tones would be tone 1 and tone 4 – i.e., the fourth row in Fig. 1A). The test sequences, which were meant to assess learning during training, could either be constructed from the same high probability transition matrix as the training stimuli (Fig. 1A) or a lower probability transition matrix (prior two tones are 65% predictive rather than 90%; Fig. 1B). Each test sequence was paired with a "non-grammatical" sequence, which was generated from a uniform transition matrix (Fig. 1C), in which the prior two tones of the sequence could not be used to predict the next tone.

2.3. Materials

The experiment was coded in jsPsych 6 (De Leeuw, 2015). A Matlab script was used to generate second-order Markov chains according to a transition probability matrix (see Fig. 1). These chains were saved as an audio file (44.1 kHz, 16-bit depth; RMS normalized). There were two frequency ranges used in the experiment. The first (hereafter referred to as "low") was identical to Durrant et al. (2011) (T_1: 261.63 Hz, T_2: 300.53 Hz, T_3: 345.22 Hz, T_4: 396.55 Hz, T_5: 455.52 Hz), whereas the second (hereafter referred to as "high") was shifted up in pitch by 50% (T_1: 392.45 Hz, T_2: 450.80 Hz, T_3: 517.83 Hz, T_4: 594.83 Hz, T_5: 683.28 Hz). Each individual tone was 200 ms in duration.

The two unfamiliar instrumental timbres were synthesized in Matlab and designed to be distinct from one another, both in terms of amplitude envelope (Fig. 2A) and harmonic spectrum (Fig. 2B). The first unfamiliar instrument (Artificial 1) had a tapered cosine envelope with a 50 ms rise and decay. The second unfamiliar instrument (Artificial 2) had a 10 ms linear rise combined with an exponential decay window. Both Artificial 1 and Artificial 2 were synthesized with seven harmonics; however, these harmonic series did not overlap (Artificial 1 consisted of harmonics 2, 3, 6, 7, 10, 11, and 13, whereas Artificial 2 consisted of harmonics 1, 4, 5, 8, 9, 12, and 14). Sample sequences generated from Artificial 1 and Artificial 2 can be heard accessed via Open Science Framework (Artificial 1: https://osf.io/67sca; Artificial 2: https://osf.io/emx8g).

The familiar instruments were sampled from a sound library associated with Reason 4 software (Propellerhead: Stockholm). We selected a piano and violin as the two familiar instruments, as these are both common instruments heard across a variety of musical genres, while still having distinct envelopes and harmonic spectra (see Fig. 2C-D; Patil, Pressnitzer, Shamma, & Elhilali, 2012). Moreover, the piano and violin have been previously shown to activate implicit representations of Western intonation (Van Hedger, Heald, Huang, Rutstein and Nusbaum, 2017). The specific piano and violin samples that we selected also shared similarities in amplitude envelope with the unfamiliar instruments. To quantify this relationship between our artificial and familiar instruments, we correlated the amplitude envelopes, extracted using a Hilbert transform, for all timbres (Artificial 1, Artificial 2, piano, violin). The strongest correlations were between Artificial 2 and piano ($r = 0.66$), followed by Artificial 1 and violin ($r = 0.60$). For comparison, all other correlations ranged between $r = 0.32$ and 0.44. Given the non-standard tuning of the constituent tones, the tuning setting was adjusted in Reason to achieve the desired frequency (e.g., 300.53 Hz could be represented as D4 + 40 cents). Individual tones were exported

A	1	2	3	4	5
11	2.5%	2.5%	2.5%	90%	2.5%
12	2.5%	2.5%	90%	2.5%	2.5%
13	2.5%	90%	2.5%	2.5%	2.5%
14	90%	2.5%	2.5%	2.5%	2.5%
15	2.5%	2.5%	2.5%	2.5%	90%
21	2.5%	2.5%	2.5%	2.5%	90%
22	2.5%	2.5%	2.5%	90%	2.5%
23	2.5%	2.5%	90%	2.5%	2.5%
24	2.5%	90%	2.5%	2.5%	2.5%
25	90%	2.5%	2.5%	2.5%	2.5%
31	2.5%	2.5%	90%	2.5%	2.5%
32	2.5%	90%	2.5%	2.5%	2.5%
33	90%	2.5%	2.5%	2.5%	2.5%
34	2.5%	2.5%	2.5%	2.5%	90%
35	2.5%	2.5%	2.5%	90%	2.5%
41	90%	2.5%	2.5%	2.5%	2.5%
42	2.5%	2.5%	2.5%	2.5%	90%
43	2.5%	2.5%	2.5%	90%	2.5%
44	2.5%	2.5%	90%	2.5%	2.5%
45	2.5%	90%	2.5%	2.5%	2.5%
51	2.5%	90%	2.5%	2.5%	2.5%
52	90%	2.5%	2.5%	2.5%	2.5%
53	2.5%	2.5%	2.5%	2.5%	90%
54	2.5%	2.5%	2.5%	90%	2.5%
55	2.5%	2.5%	90%	2.5%	2.5%

B	1	2	3	4	5
11	8.75%	8.75%	8.75%	65%	8.75%
12	8.75%	8.75%	65%	8.75%	8.75%
13	8.75%	65%	8.75%	8.75%	8.75%
14	65%	8.75%	8.75%	8.75%	8.75%
15	8.75%	8.75%	8.75%	8.75%	65%
21	8.75%	8.75%	8.75%	8.75%	65%
22	8.75%	8.75%	8.75%	65%	8.75%
23	8.75%	8.75%	65%	8.75%	8.75%
24	8.75%	65%	8.75%	8.75%	8.75%
25	65%	8.75%	8.75%	8.75%	8.75%
31	8.75%	8.75%	65%	8.75%	8.75%
32	8.75%	65%	8.75%	8.75%	8.75%
33	65%	8.75%	8.75%	8.75%	8.75%
34	8.75%	8.75%	8.75%	8.75%	65%
35	8.75%	8.75%	8.75%	65%	8.75%
41	65%	8.75%	8.75%	8.75%	8.75%
42	8.75%	8.75%	8.75%	8.75%	65%
43	8.75%	8.75%	8.75%	65%	8.75%
44	8.75%	8.75%	65%	8.75%	8.75%
45	8.75%	65%	8.75%	8.75%	8.75%
51	8.75%	65%	8.75%	8.75%	8.75%
52	65%	8.75%	8.75%	8.75%	8.75%
53	8.75%	8.75%	8.75%	8.75%	65%
54	8.75%	8.75%	8.75%	65%	8.75%
55	8.75%	8.75%	65%	8.75%	8.75%

C	1	2	3	4	5
11	20%	20%	20%	20%	20%
12	20%	20%	20%	20%	20%
13	20%	20%	20%	20%	20%
14	20%	20%	20%	20%	20%
15	20%	20%	20%	20%	20%
21	20%	20%	20%	20%	20%
22	20%	20%	20%	20%	20%
23	20%	20%	20%	20%	20%
24	20%	20%	20%	20%	20%
25	20%	20%	20%	20%	20%
31	20%	20%	20%	20%	20%
32	20%	20%	20%	20%	20%
33	20%	20%	20%	20%	20%
34	20%	20%	20%	20%	20%
35	20%	20%	20%	20%	20%
41	20%	20%	20%	20%	20%
42	20%	20%	20%	20%	20%
43	20%	20%	20%	20%	20%
44	20%	20%	20%	20%	20%
45	20%	20%	20%	20%	20%
51	20%	20%	20%	20%	20%
52	20%	20%	20%	20%	20%
53	20%	20%	20%	20%	20%
54	20%	20%	20%	20%	20%
55	20%	20%	20%	20%	20%

Fig. 1. Transition matrices underlying the construction of the tone sequences for high grammatical (A), low grammatical (B), and ungrammatical (uniform) sequences (C).
Note: The row indexes the last two tones that have occurred, the column indexes the next tone that could occur, and the value gives the probability of this transition. For example, if listeners heard Tone 5 followed by Tone 5 (final row), there would be a 90% probability of hearing Tone 3 in the high grammatical sequences (A), a 65% probability of hearing Tone 3 in the low grammatical sequences (B), and a 20% probability of hearing Tone 3 in the ungrammatical (uniform) sequences (C). Training sequences always adhered to the high grammatical probabilities in panel A.

as audio files and then imported into Matlab to generate the test and training sequences.

The training sequence contained 2000 tones, corresponding to a duration of 6 min and 40 s. This sequence was subdivided into smaller sequences consisting of 200, 400, 600, and 800 tones, which allowed us to incorporate a behavioural task designed to ensure task compliance and monitor participants' attention (see Section 2.5 for details). The sequences used for testing consisted of 22 notes (4.4 s in duration) which resulted in 20 transitions that could either adhere to or violate the transition matrix. The *high grammatical* test sequences consisted of 17, 18, or 19 high-probability transitions, generated with the transition matrix in Fig. 1A. The *low grammatical* test sequences consisted of 12, 13, or 14 high-probability transitions, generated with the transition matrix in Fig. 1B. The inclusion of both high and low grammatical trials parallels Durrant et al. (2011) and increases the sensitivity of the test by probing performance across a range of difficulty levels. The *ungrammatical* stimuli were generated with a uniform transition probability matrix (i.e., there was a 20% probability of hearing any of the five tones), represented in Fig. 1C. As such, although some of the transitions in the ungrammatical sequences adhered to the second-order Markov transitions displayed in Figs. 1A-B, these occurred much less frequently than in the high and low grammatical sequences, and the mean number of "grammatical" transitions was constrained to be 20% (M = 4.00, SD = 1.68, range: 0 to 8). Fig. 3B provides samples of each type of grammatical sequence (high, low, and ungrammatical) generated by the transition matrices from Fig. 1. The training sequences always adhered to the high grammatical transition matrix. During testing, either a high or low grammatical sequence was always paired with an ungrammatical sequence.

The high, low, and ungrammatical sequences were well matched on general contour and interval measures (see Supplementary Material 1: *Analysis of Contours and Intervals in Test Sequences*). In particular, the three stimulus types did not differ in terms of (1) the number of tone repeats, (2) the total number of ascending intervals, (3) the total number of descending intervals, (4) mean interval size, (5) standard deviation of the interval size, (6) the distribution of specific intervals, or (7) first-order (i.e., tone-to-tone) transitions, represented in terms of general contour and specific interval size. The fact that the three stimulus types were matched on these contour and interval attributes means that participants could not rely on more local cues (e.g., the number of times a

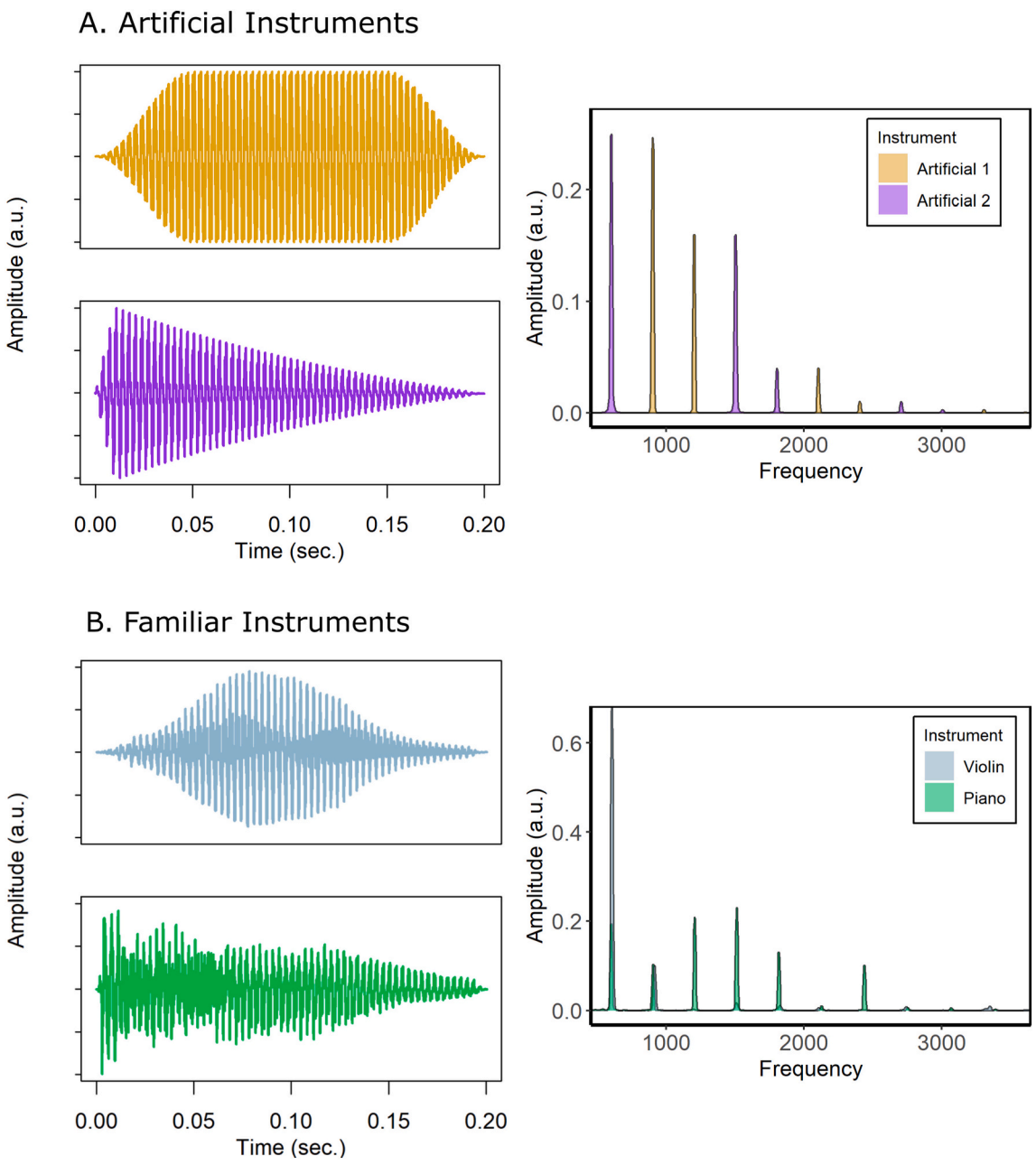

Fig. 2. Visualization of the auditory stimuli used in the experiment.
Note: (A) Amplitude envelope (left) and harmonic spectrum (right) of the artificial tones. (B) Amplitude envelope (left) and harmonic spectrum (right) of the familiar tones.

tone repeated, the relative distribution of large and small interval changes) to accurately judge whether a sequence was grammatical or not. This is important to note given that prior grammar sequence learning studies have discussed that patterns of repetition may alert individuals to the grammar category of a test sequence (Brooks & Vokey, 1991; Tunney & Altmann, 2001), including within atonal auditory sequences (Endress, Dehaene-Lambertz, & Mehler, 2007).

2.4. Procedure

After providing informed consent, participants completed an auditory calibration and headphone assessment. First, participants were presented with a 30-s pink noise, root-mean-square normalized to the same level as the auditory sequences and were asked to adjust their computer's volume to a comfortable listening level. Next, participants completed a short headphone assessment in which, on each trial, they judged which of three sounds was quietest (see Woods, Siegel, Traer, & McDermott, 2017). The sounds were designed such that the loudness judgments should be easy when wearing headphones but difficult if participants were listening over standard computer speakers. There were six trials in total.

Participants then completed the main statistical learning task (Fig. 3). Participants were first introduced to the training component of the experiment. The instructions stated that participants would hear a long sequence of tones, which may sound like an unfamiliar melody. The training sequence was generated from the high grammatical transition matrix shown in Fig. 1A, meaning the prior two tones of the sequence would result in the "expected" transition 90% of the time. Participants were instructed to monitor this sequence for interruptions – i.e., periods of silence that could last up to ten seconds. Whenever participants heard

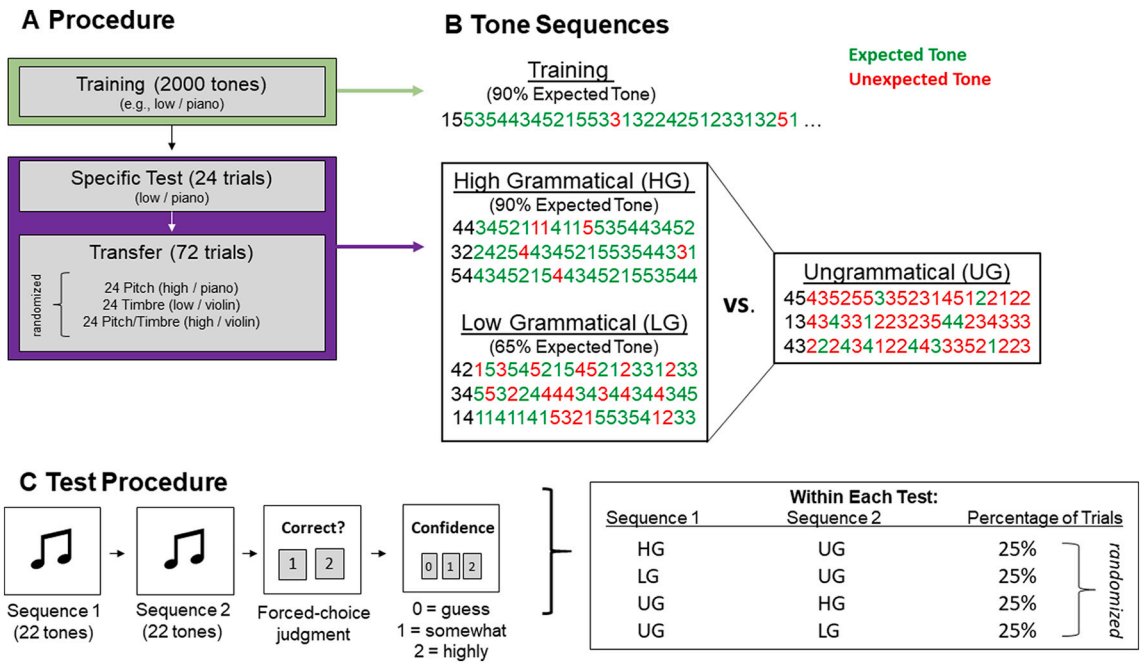

Fig. 3. Overview of the training and testing procedure.

Note: Panel A outlines the flow of the experiment, in which participants first completed a training block consisting of 2000 tones that adhered to the high probability transition matrix. Following this training, participants completed two assessments of learning. The Specific Test used short test sequences that were identical in pitch and timbre to the training sequences. The Transfer Test used short test sequences that could differ in either pitch or timbre from the training sequence. Panel B provides samples of the different categories of test sequences. High grammatical (HG) sequences contained an average of 90% expected transitions, low grammatical (LG) sequences contained an average of 65% expected transitions, and ungrammatical (UG) sequences contained an average of 20% expected transitions. Expected transitions are printed in green. Panel C outlines the general procedure for the Specific and Transfer Tests. Each trial consisted of a forced-choice judgment, in which participants had to judge whether the first or second sequence sounded more like the training sequence. Participants also provided a confidence rating in their answer.

such an interruption, they were instructed to press the spacebar as quickly as possible. If participants did not press the spacebar within ten seconds, the experiment advanced automatically, and the trial was marked as a timeout. There were four "interruption-detection" trials, corresponding to the breaks between the subdivided audio files for training (200, 400, 600, and 800 tones). As such, the shortest span of time between responses was 40 s (for the 200-tone audio file) and the longest span of time between responses was 2 min 40 s (for the 800-tone audio file). The four audio files were presented randomly. This interruption-detection task was implemented given our Internet-based sample, which precluded direct participant monitoring. Performance on this interruption-detection task was used to cull participants from the primary analyses (See Section 2.5). Within each instrument training group (Artificial, Familiar) the instrument (e.g., Artificial 1 vs. Artificial 2 or piano vs. violin) and pitch level (low vs. high) for the training sequence was randomly assigned across participants. Thus, there were eight total between-participant combinations of the training sounds (Artificial Instrument condition: Artificial 1/high, Artificial 1/low, Artificial 2/high, Artificial 2/low, Familiar Instrument condition: piano/high, piano/low, violin/high, violin/low).

Following training, participants completed the Specific Test, which used the same instrument and pitch level as the training sequence, and a two-alternative forced-choice (2AFC) procedure (Fig. 3C). On each trial, participants heard two shorter sequences of tones, with one of the two sequences being grammatical (either high or low), and the other sequence being ungrammatical After hearing both sequences, participants indicated which alternative (first or second) sounded more like the training sequence. Participants then rated their confidence on a three-point scale (*complete guess*, *somewhat confident*, and *extremely confident*). Although the resolution of the confidence judgments was coarse, it provided a means of assessing whether participants had accurate meta-cognition of their judgments (e.g., whether "complete guess" trials were indeed at chance or whether accuracy increased with confidence). There were 24 trials in the Specific Test (12 trials in which the correct answer was a high grammatical sequence and 12 trials in which the correct answer was a low grammatical sequence), with the grammatical sequence as the first alternative 50% of the time. The Specific Test included a simple auditory attention check (in which an audio file played and notified participants to click on one of three labeled buttons presented on the screen).

Participants then completed the Transfer Test, which was also a 2AFC procedure but was three times as long, as it consisted of three subtests, with trials in which both alternatives were: (1) sequences that differed in timbre but had the same auditory frequency range as training (Transfer/Timbre Test); (2) sequences that had the same timbre as the training but had an auditory frequency range shifted 50% relative to training (Transfer/Pitch Test), and (3) sequences that differed in timbre and also had a 50% shifted auditory frequency range compared to the training (Transfer/Timbre-Pitch Test). Trials from these three conditions were presented in random order within the Transfer Test. Participants were explicitly instructed that the sequences could differ in timbre or auditory frequency (framed to the participants in terms of instrument and pitch height, respectively) compared to the previous parts of the experiment. However, the instructions also emphasized that, despite these differences, one of the two alternatives in each trial would still sound more like the training sequence. Each of the three conditions was tested with 24 trials (12 trials in which the correct answer was a high grammatical sequence and 12 trials in which the correct answer was a low grammatical sequence), with the grammatical sequence in the first alternative 50% of the time. The Transfer Test also included a simple auditory attention check (in which an audio file played and notified participants to click on one of three labeled buttons presented on the screen).

Following the Transfer Test, participants completed a basic

demographic, language experience, music experience, and hearing assessment questionnaire. After this questionnaire, participants were redirected to Qualtrics, in which they completed a password-protected short-form version of the Raven's Advanced Progressive Matrices (Arthur & Day, 1994). Participants were given 20 min to solve 12 items from the Advanced Matrices Set.

2.4.1. Follow-up sessions

Participants were invited to complete two follow-up sessions to assess retention of learning. The first follow-up (Session 1) was available between 24 and 48 h after the initial session (Session 1), and the second follow-up (Session 3) was available one week after Session 1. The majority of participants in both the Artificial (41 of 51) and Familiar (46 of 53) Instrument Training Groups completed Session 3 the day it became available (i.e., seven days post-training; $M \pm SD$: 7.27 ± 0.77 days, range of 7 to 12 days). The procedure was similar to Session 1, with the key difference being that the two follow-up sessions did not include training. After providing informed consent, participants completed a basic volume adjustment and headphone assessment. Following this auditory calibration, participants completed the Specific Test (which used the same instrument/pitch combination that the participant experienced in training of Session 1). After the Specific Test, participants completed the Transfer Tests (Transfer/Timbre, Transfer/Pitch, Transfer/Pitch-Timbre). The tests in Sessions 2 and 3 were identical in structure to the ones administered in Session 1, although the specific note sequences were always novel, as we did not want memory for specific exemplars to influence performance (cf. Agus, Thorpe, & Pressnitzer, 2010). In Session 2, following the Transfer Tests, participants completed a questionnaire assessing their sleep duration and quality the previous night (i.e., to confirm that participants slept). There were no additional assessments following the Transfer Tests in Session 3. Data across sessions were linked via participants' Amazon MTurk Worker IDs. Participants were compensated at the end of each session.

2.5. Participant culling

The data were subjected to three culling considerations prior to full analysis. Failure to pass these considerations resulted in the participant being excluded from the analyses. The first consideration was that participants detected at least three of the four interruptions during the training sequence (i.e., achieved at least 75% accuracy). Ten participants in the Artificial Instrument condition and 14 participants in the Familiar Instrument condition were discarded at this stage, leaving 90 analyzable participants in the Artificial Instrument Condition and 84 analyzable participants in the Familiar Instrument Condition. The majority of the remaining participants performed perfectly on the interruption-detection task (89 of 90 Artificial Instrument participants, 80 of 84 Familiar Instrument participants). The second consideration was that participants correctly answered at least one of the two auditory attention checks presented during the Specific and Transfer Tests. No additional participants were discarded based on this threshold. Overall, performance on the auditory attention checks was high (Artificial Instrument: 88 of 90 analyzable participants correctly answered both checks; Familiar Instrument: 80 of 84 analyzable participants correctly answered both checks).

The third consideration was that participants completed the entire experiment – i.e., all three experimental sessions. Of the 174 participants who completed Session 1 and passed the attention checks, 132 participants completed Session 2 (Artificial: $n = 73$, Familiar: $n = 59$). A total of 104 participants successfully completed Session 3 (Artificial: $n = 51$, Familiar: $n = 53$). Given that our hypotheses related to how learned tonal sequences are remembered, only participants who completed all three sessions were considered in the primary analyses. Table 1 provides a comparison of participants who completed all testing sessions in both Instrument Training Groups.

Table 1
Comparison of participants across instrument conditions.

Measure	Artificial (n = 51)	Familiar (n = 53)	BF_{01}
Age (years)	37.16 ± 10.90	38.04 ± 11.20	4.48
Gender (prop. female)	0.47 ± 0.50	0.51 ± 0.50	4.50
Bachelor's Degree (prop. yes)	0.43 ± 0.50	0.53 ± 0.50	3.13
Num. Lang. (familiar with)	1.47 ± 0.70	1.53 ± 0.85	4.53
Musician (prop. yes)	0.61 ± 0.49	0.51 ± 0.50	3.08
Hearing Issues (yes)	0.04 ± 0.20	0.15 ± 0.36	0.90
Raven's Matrices (max. 12)	5.59 ± 2.55	5.15 ± 2.81	3.55

Note: The BF_{01} reflects the relative evidence in favor of the null hypothesis. For example, comparing the mean ages of the two groups suggests that the null hypothesis is 4.48 times more likely than the alternative hypothesis that the groups differ in age. These analyses suggest that the two groups were well-matched on all measured variables. Artificial: Artificial Instrument Condition, Familiar: Familiar Instrument Condition.

2.6. Data analysis

All analyses used generalized linear mixed-effects models (GLMMs) with a binomial link, given that performance on each trial was binary (i.e., correct/incorrect). The GLMMs were run in R (R Core Team, 2016) using the *lme4* package (Bates, Mächler, Bolker, & Walker, 2015). In Section 3.1, we assessed whether the proportion of correct responses was above the chance estimate (i.e., 0.50). This analysis used an intercept-only GLMM. This intercept-only model additionally modeled random intercepts for participant and sequence (both grammatical and ungrammatical). In Section 3.2, we assessed how performance accuracy varied as a function of our experimental factors (Instrument Training Group, Session, Difficulty, and Test). In this GLMM, we assessed the main effect of Instrument Training Group on performance, as well as the interactions of Instrument Training Group with Session (1,2,3), Difficulty (high grammaticality, low grammaticality), and Test (Specific, Transfer/Pitch, Transfer/Timbre, Transfer/Pitch-Timbre). Session was treated as an ordered factor. The Specific Test was used as the reference category. This GLMM additionally modeled random intercepts for participant and sequence (both grammatical and ungrammatical).

In Section 3.3, we assessed how participant accuracy related to confidence judgments, to determine whether participants in the Familiar Instrument Training Group might exhibit overconfidence relative to Artificial Instrument participants, presumably because confidence could be influenced by instrumental familiarity (in addition to perceived sequence familiarity). The GLMM approach was particularly advantageous in this case because it could account for the uneven observations of confidence ratings across participants. Confidence judgments were treated as an ordered factor and were modeled to interact with Instrument Training Group. The GLMM also included Session, Difficulty, and Test. Participant and sequence intercepts were included as random effects.

In Section 3.4 we modeled performance accuracy as a function of the specific intervallic features of each sequence – in other words, features related to the pattern of rises and falls in the melodic sequences. Given that the high grammatical, low grammatical, and ungrammatical sequences did not differ on a variety of contour and interval measures, including first-order note-to-note transitions (see Table S2), relying on this adjacent interval information would not help a listener make an accurate judgment. However, if familiar instruments biased listeners to perceive and process the novel sequences more musically, then we would expect a stronger association between intervallic information and grammaticality judgments, as adjacent changes in relative pitch are critical for melodic understanding in Western music (e.g., Attneave & Olson, 1971). To assess how particular features of the test sequences related to performance, we created GLMMs for a variety of interval measures (see Table 2). In each model, accuracy (correct, incorrect) was the dependent variable. Each model contained interval information for both the grammatical and the ungrammatical sequence for each trial (e.

Table 2
Summary of interval model terms predicting performance accuracy.

Interval measure	Grammatical sequence	Ungrammatical sequence	Instrument x grammatical sequence	Instrument x ungrammatical sequence
Tone Repeats (#)	B = −0.028 (0.014) +	**B = 0.024 (0.010)** *	B = 0.010 (0.020)	B = −0.005 (0.013)
Ascending Contours (#)	B = −0.010 (0.014)	B = −0.023 (0.012) +	B = 0.013 (0.020)	B = 0.007 (0.017)
Descending Contours (#)	**B = 0.052 (0.017)** **	B = −0.020 (0.013)	B = −0.033 (0.023)	B = 0.000 (0.019)
Mean Interval Size	B = 0.041 (0.084)	**B = −0.29 (0.064)** ***	**B = 0.283 (0.116)** *	B = −0.064 (0.089)
St. Dev. of Interval Size	B = 0.029 (0.125)	**B = −0.211 (0.101)** *	B = 0.275 (0.172)	B = −0.088 (0.140)
± 2/± 3 Intervals	B = −0.005 (0.030)	**B = −0.113 (0.032)** ***	**B = 0.082 (0.041)** *	B = −0.02 (0.044)

Note: Standard errors are represented in parentheses. + p < .10 * p < .05 ** p < .01 *** p < .001.

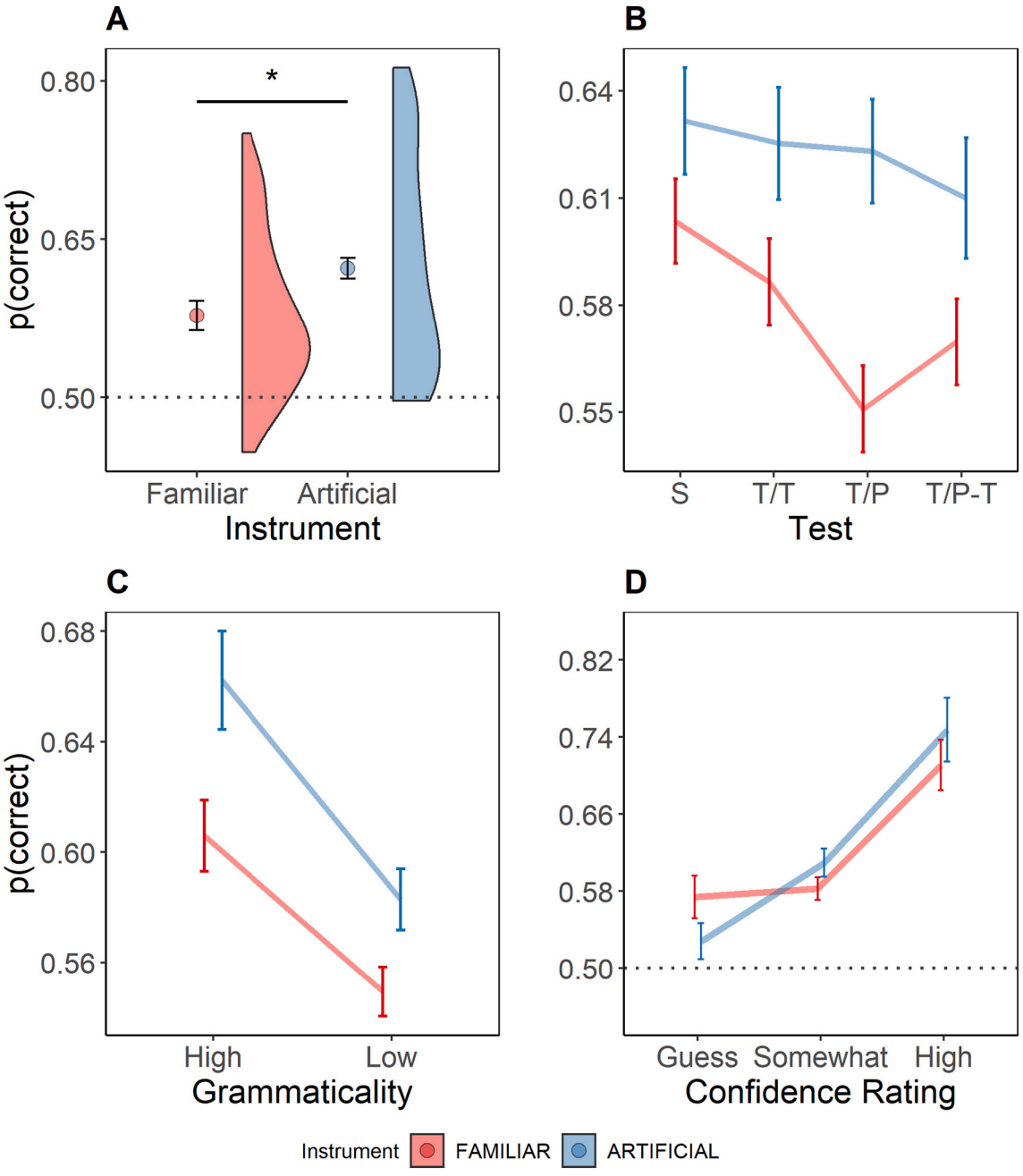

Fig. 4. Instrument-related effects on experiment factors.
Note: (A) Main effect of Instrument training group, depicted by summary statistics (dot and error bar, left) and distributional representation (right). (B) Interaction of Test and Instrument Training Group. (C) Interaction of Instrument and Difficulty. (D) Interaction of Instrument and Confidence Rating. Error bars represent ± 1 standard error of the mean. Across panels, Artificial Instrument is plotted in blue and Familiar Instrument is plotted in red. S: Specific Test, T/T: Transfer/Timbre Test, T/P: Transfer/Pitch Test, T/P-T: Transfer/Pitch-Timbre Test, * p < .05.

g., the number of repeats contained within the grammatical and ungrammatical sequence). Instrument Training Group was included as a main effect and was also allowed to interact with the interval measures. We additionally included Session, Difficulty, and Test in each model. Only participant intercepts were included as random effects, as including sequence intercepts resulted in nonconvergence of the models.

3. Results

3.1. Testing performance against chance

The grand mean of accuracy across all factors was 60.0%, which was well above the chance estimate (B = 0.43, $z = 10.11$, $p < .001$). This evidence of successful learning conceptually replicates previous work (Durrant et al., 2011) using an Internet-based sample. Participants in both the Artificial (B = 0.54, $z = 8.05$, $p < .001$) and Familiar (B = 0.32, $z = 7.34$, $p < .001$) Instrument Training Groups were independently above chance.

3.2. Modeling performance in terms of experimental factors

There was a significant main effect of Instrument Training Group (B = −0.19, $z = 2.14$, $p = .033$), with Familiar Instrument participants performing more poorly than Artificial Instrument participants (57.8% vs. 62.2%; Fig. 4A). Furthermore, relative to the Specific Test, this attenuation in performance for the Familiar Instrument participants was particularly pronounced in the Transfer/Pitch Test, as evidenced by a significant interaction between Instrument and Transfer/Pitch (B = −0.19, $z = -2.70$, $p = .007$; Fig. 4B).

There was a large main effect of Difficulty (B = −0.36, $z = -8.92$, $p < .001$), with participants performing better on high grammatical trials (66.3%) compared to low grammatical trials (56.6%). Instrument Training Group additionally interacted with Difficulty (B = 0.12, $z = 2.38$, $p = .017$; Fig. 4C). This interaction was characterized by pronounced accuracy differences between Artificial and Familiar Instrument participants for high grammatical trials (66.2% versus 60.6%, respectively), as opposed to more comparable, attenuated performance for low grammar trials (58.3% versus 54.9%, respectively).

Instrument Training Group did not significantly interact with Session (B = −0.28, $z = -0.68$, $p = .499$), suggesting that the attenuation in performance in the Familiar Instrument participants was not significantly modulated by timescale. There was additionally no significant main effect of Session (B = −0.01, $z = -0.28$, $p = .780$), indicating the performance was maintained across the three time points. We observed a marginal effect of the Transfer/Pitch Test (B = −0.09, $z = -1.83$, $p = .068$), meaning performance in this Transfer Test was attenuated relative to the Specific Test (see Fig. 4B). This marginal main effect was primarily driven by the Familiar Instrument participants, as evidenced by the interaction between Instrument Training Group and the Transfer/Pitch Test described previously. No other terms in the model were significant.

3.3. Confidence ratings

Participants in both Instrument Training Groups showed similar distributions of confidence ratings (Artificial Instrument: Guess = 22%, Somewhat = 59%, High = 18%; Familiar Instrument: Guess = 24%, Somewhat = 58%, High = 18%), suggesting that participants in both groups used the confidence ratings in a similar manner (χ^2 (2) < 1). Yet, performance on guess trials was significantly above chance for participants in both the Familiar (B = 0.14, $z = 3.12$, $p = .002$) and Artificial (B = 0.21, $z = 4.12$, $p < .001$) Instrument Training Groups, suggesting some dissociation between metacognition and accuracy.

In our model, we observed a highly significant effect of confidence ratings. This effect of confidence ratings was significant using both a linear fit (B = 0.82, $z = 17.28$, $p < .001$) and quadratic fit (B = 0.22, $z = 6.72$, $p < .001$), reflecting the fact that participants tended to have higher accuracy as a function of higher confidence ratings, with a particular difference between "somewhat" and "high" confidence answers. Importantly, both the linear and quadratic effects of confidence rating significantly interacted with Instrument Training Group (linear: B = −0.35, $z = 5.40$, $p < .001$; quadratic: B = −0.16, $z = -3.53$, $p < .001$), with Familiar Instrument participants showing *attenuated* linear and quadratic associations of confidence with accuracy (Fig. 4D). The significant main effects of Instrument Training Group (B = −0.28, $z = -3.54$, $p < .001$) and Difficulty (B = −0.26, $z = -8.34$, $p < .001$) reported in Section 3.2 were also observed in this model.

3.4. Predicting accuracy from intervals

Two main findings emerged from the models predicting accuracy from contour and intervallic features of the test sequences. First, interval information in the ungrammatical sequences predicted performance accuracy more strongly than interval information in the grammatical sequences. Of the six interval measures described in Table 2, five were significantly or marginally associated with accuracy for ungrammatical sequences. Essentially, this means that despite the random construction of ungrammatical trials (Fig. 1C), participants were more likely to judge these sequences as grammatical if they contained fewer note repeats and had relatively larger and more variable intervals. This suggests that participants were basing their decisions of grammaticality on the (perceived) structure of the ungrammatical sequences (e.g., Brooks & Vokey, 1991). These effects in the ungrammatical sequences, however, did not interact with Instrument Training Group.

Second, we found that Instrument Training Group interacted with mean interval size in the grammatical sequences, with Familiar Instrument participants' accuracy positively associated with *larger* mean interval sizes. This finding may be explained by the nature of the tonal system used in the present study. Specifically, by dividing the octave into five equal steps, intervals of ±2 tones (e.g., Tone 1 to Tone 3) and ± 3 tones (e.g., Tone 1 to Tone 4) correspond to changes of 480 cents and 720 cents, respectively. These intervals are therefore quite close (±20 cents, or ~1.2%) to perfect fourth (500 cents) and perfect fifth (700 cents) intervals in Western music. Perfect fourths and fifths have a prioritized status in Western tonal music, as they are rated as highly consonant (Krumhansl, 1990) and represent common harmonic transitions (Krumhansl, Bharucha, & Castellano, 1982).

Given this incidental relationship between ±2/± 3 intervals in the present tonal system and perfect fourths/fifths in Western music, we calculated the mean number of ±2 and ±3 intervals in each test sequence to assess whether the prevalence of these intervals was associated with accuracy (Table 2, final row). We found a main effect of the number of ±2 and ±3 intervals within the ungrammatical sequences on accuracy, with a greater number of these intervals being associated with lower accuracy (i.e., a greater likelihood of judging the ungrammatical sequence as grammatical). Critically, we also observed a significant interaction between Instrument Training Group and the number of ±2 and ±3 intervals within the grammatical sequences, with Familiar Instrument participants showing a strong, positive relationship between the number of these intervals and accuracy and the Artificial Instrument participants showing no relationship between these intervals and accuracy (Fig. 5).

4. Discussion

The present results suggest that instrument familiarity influences the statistical learning of novel tonal sequences. We found that the same tone-learning task – at least from an abstract, melodic perspective – results in different patterns of performance depending on whether the timbre is that of a familiar instrument or not. Compared to participants who trained and tested with artificial timbres, participants who were trained and tested with familiar instruments displayed (1) attenuated

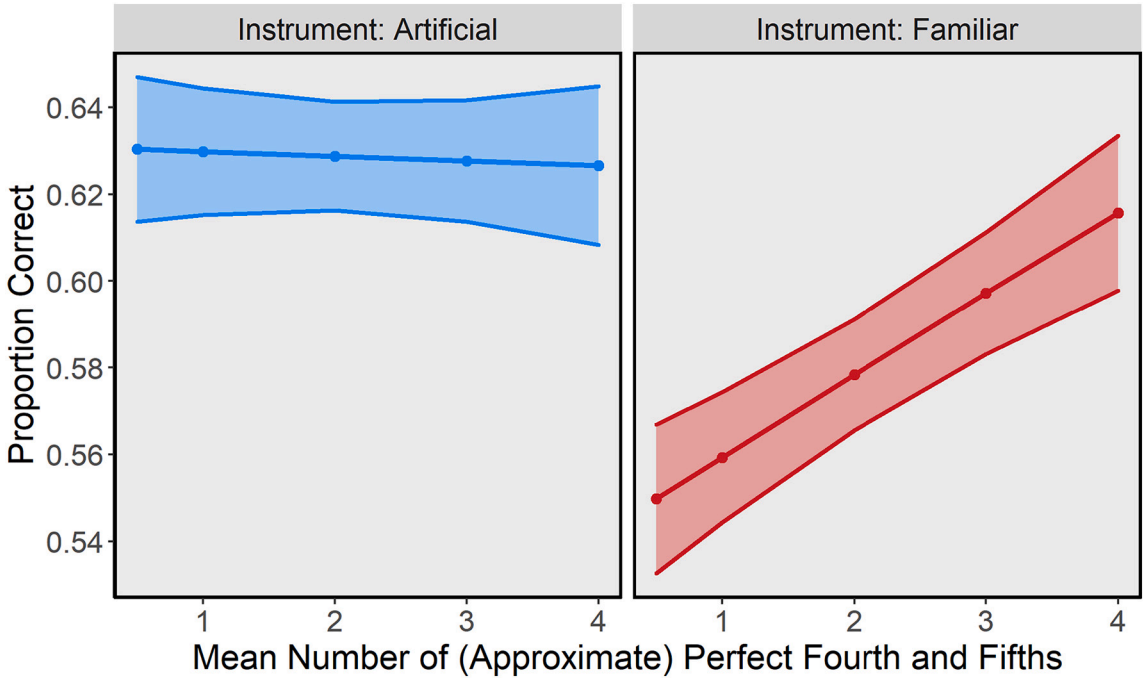

Fig. 5. Interaction of approximate perfect fourths and fifths and instrument familiarity in predicting accuracy.
Note: Plots represent the mean number of approximate perfect fourths and fifths for grammatical sequences. The mean number of approximate perfect fourth and fifths was calculated by averaging the number of − 2, +2, −3, and + 3 intervals in each sequence, as ± 2 intervals correspond to a change of 480 cents and ± 3 intervals correspond to a change of 720 cents. Both intervals are within 20 cents (1.2%) of perfect fourths (500 cents) and perfect fifths (700 cents) in Western music. Ribbons represent ± 1 standard error of the mean.

learning, (2) a larger dissociation between confidence and accuracy, and (3) a greater reliance on "musical" intervals in judging grammaticality, despite these cues being orthogonal to the present learning task. These findings not only extend the notion of knowledge entrenchment to a non-linguistic domain, but they also highlight that the influence of prior knowledge on statistical learning performance is sensitive to context and is shaped by stimulus features that are incidental to the assigned learning task.

We deviated from prior investigations of knowledge entrenchment (Siegelman et al., 2018) both in terms of domain (musical versus linguistic knowledge) and in terms of paradigm (probabilistic learning of non-repeating sequences versus learning of specific trisyllabic "words"). Despite these differences, the results from the present experiment are consistent with a general framework of knowledge entrenchment across several measures. First, overall performance was poorer for familiar instruments, particularly for sounds that differed from the training sounds in pitch (i.e., Transfer/Pitch trials). This result supports the idea that familiar instruments more strongly evoked prior musical categories (e.g., representations of melodic sequences that pianos and violins typically play), with this activation of prior musical knowledge interfering with learning the present tonal grammar. Second, participants listening to familiar instruments had a significantly attenuated relationship between confidence and accuracy compared to participants listening to artificial instruments. This suggests that participants listening to familiar instruments had less meta-awareness of their knowledge – particularly when claiming "high" levels of confidence, possibly because confidence ratings were being influenced by additional factors such as instrument familiarity (cf. Busey, Loftus, & Loftus, 2000; Chua et al., 2012). Third, the prevalence of particular "musical" intervals biased participants' judgments more in the familiar instrument compared to the artificial intrument condition, *despite the fact that the prevalence of these intervals was not associated with the current tone grammar*.

While these results support the general principle of knowledge entrenchment in a novel domain (music), there are also a number of commonalities between the present results and previous reports of entrenchment effects in the domain of language. First, the overall attenuation in performance has been reported in a prior linguistic statistical learning study (e.g., Finn & Hudson Kam, 2008). This study reported attenuated learning when the Finn and Hudson Kam, 2008 report attenuated learning when the statistical sequences contain phonotactic violations (e.g., beginning a "word" with *tl*). In this sense, our paradigm is closely aligned with this approach taken by Finn & Hudson Kam, 2008, as both our individual notes and the intervals between notes violated the conventional rules of Western music. As such, it is possible that an increased familiarity with the constituent items in language-based statistical learning tasks scaffolds performance, meaning entrenchment is not observed as a simple main effect of performance when the individual items, as well as the transitions between items, are plausible in the particular domain. Second, the dissociation of confidence and accuracy conceptually aligns with language-based investigations of entrenchment, although to our knowledge these prior investigations have not explicitly examined participant confidence in learning as a measure of entrenchment. Third, our result showing an increased likelihood of judging a sequence as grammatical if it contains more "musical" intervals for familiar instruments is perhaps the most aligned with language-based reports of entrenchment. In both cases, performance at the individual trial level can be (somewhat) predicted by the extent to which overlap occurs between the experimental stimuli and prior knowledge. For example, Siegelman et al. (2018) reported that recognition performance correlates with the extent to which the target and foil stimuli resemble words in participants' native language. Similarly, in the present experiment, we demonstrate that the incidental prevalence of "musical" intervals in target sequences influences participants' judgments of grammaticality. Additionally, it is important to note that the present approach differs from language-based entrenchment studies in the sense that we demonstrate that patterns of entrenchment can change based on manipulating a more surface-level attribute (timbre) that should be irrelevant to learning the statistical structure of the sequences. In a language-based paradigm, this might be akin to

generating syllables and being able to manipulate the extent to which knowledge entrenchment is observed based on whether the sounds are interpreted as language by the participants.

The fact that the manipulation of instrument was incidental to the learning task and not explicitly mentioned to participants makes the present results particularly compelling as evidence for the dynamic nature of knowledge entrenchment. At the same time, the present results also align with previous demonstrations that instrumental timbre is an important contextual cue in the activation of prior knowledge for both musical experts and novices. Many individuals with absolute pitch show some type of performance cost associated with switching between instruments (Van Hedger, Heald, & Nusbaum, 2015) or categorizing notes that are played with unfamiliar timbres in musical contexts (e.g., Lockhead & Byrd, 1981). Additionally, most individuals, regardless of explicit musical training, have developed implicit associations of instrumental timbre and specific pitches (Van Hedger, Heald, Huang, Rutstein, & Nusbaum, 2017). In this study, listeners judged whether isolated notes were either "in-tune" or "out-of-tune" (according to conventional Western tuning standards). Although participants were above chance in making these judgments for familiar instruments (piano and violin), performance was at chance for computer-generated complex tones that are not commonly heard in musical contexts. Taken together, the results from both the explicit and implicit absolute pitch literature suggest that listeners have developed source-specific representations that can differentially engage prior knowledge. More broadly, this suggests that listeners are constantly extracting information from different sources in the environment in order to optimize processing and prediction.

The incidental use of instrumental timbre in the current experiments also suggests that prior knowledge might not need to be explicitly or intentionally activated in order to observe effects of entrenchment. That being said, we did not assess participants' awareness of this manipulation (e.g., their explicit recognition of the instruments as piano and violin), and thus it is unclear whether the observed entrenchment in the present experiments operates explicitly or more implicitly (i.e., occurring independently of participants' explicit recognition of the familiar instruments). Given that there are both explicit and implicit contributions to statistical learning (Batterink, Reber, Neville, & Paller, 2015), and to auditory processing (Holmes et al., 2018), exploring the role of intention and awareness will be an important future direction in further specifying the processes through which prior experiences can influence statistical learning. Additionally, it should be noted that, even though familiar instrument performance was attenuated relative to artificial instruments, it was still independently above chance. Thus, although prior knowledge disrupted statistical learning of novel (incompatible) tone sequences, learning was not completely disrupted. This is not surprising, as learning requires both the ability to integrate new knowledge, as well as stability in order to prevent the forgetting of previous knowledge (e.g., Mcclelland, Mcnaughton, & Reilly, 1995). This consideration also raises an interesting theoretical possibility of whether the influence of prior knowledge can be sufficiently strong in some contexts to entirely disrupt learning, as has been found for phonotactic knowledge in linguistic statistical learning (Finn & Hudson Kam, 2008). This represents a potential avenue for future research to explore, perhaps through systematically manipulating multiple features (i.e., beyond instrument) to assess if greater degrees of entrenchment lead to graded performance declines.

The present results did not support different trajectories of long-term memory retention for familiar versus artificial instrument sequences in the days following learning. Situated in a larger context of statistical learning and sleep-dependent memory consolidation, these results do not support the claim that statistical learning performance on this probabilistic sequence learning task is *enhanced* relative to baseline following a delay period containing sleep (e.g., Durrant et al., 2011). However, the fact that statistical learning performance for the Transfer Test trials was stable for at least one week after training across all participants, deviating by less than one percentage point (Session 1: 58.92%; Session 2: 59.57%; Session 3: 59.63%), is consistent with previous research showing that statistical learning produces stable long-term knowledge. For example, memory for statistical learning appears robust over the span of a day (Kim, Seitz, Feenstra, & Shams, 2009) and even up to one year after initial learning (Kóbor, Janacsek, Takács, & Nemeth, 2017). In addition, stimulus-specific attributes (such as instrumental timbre) clearly have lasting effects on learning and retention, beyond the short-lived effects on overall performance (cf. Schellenberg & Habashi, 2015).

The present results do not inherently challenge conclusions stemming from prior investigations of implicit tone sequence learning, as most studies have used sine tones (e.g., Creel, Newport, & Aslin, 2004; Durrant et al., 2011; Durrant, Cairney, & Lewis, 2013; Furl et al., 2011; Loui, 2012; Loui et al., 2010; Loui & Wessel, 2008), which are uncommon in musical contexts and would not have strong associations with implicitly acquired musical knowledge (e.g., Van Hedger, Heald, Huang, Rutstein and Nusbaum, 2017). However, the selection of timbre in these studies is rarely justified, especially compared to the thought that is put into selecting unfamiliar pitch intervals (e.g., from the Bohlen-Pierce scale; Mathews, Pierce, Reeves, & Roberts, 1988). Thus, one practical recommendation from the present findings is for researchers to carefully consider how stimuli used in statistical learning paradigms could engage prior knowledge, taking into account both "higher-order" elements (such as how tone or syllable sequences relate to known musical or linguistic patterns) as well as more perceptually oriented attributes (such as timbre).

4.1. Possible mechanisms underlying knowledge entrenchment

The present results are broadly consistent with the notion of knowledge entrenchment (Siegelman & Frost, 2015), but the precise cognitive mechanisms underlying these effects remain to be clarified. One possibility is that prior knowledge shapes the perception, encoding and/or long-term memory representations of the *individual items* in the present tonal system. In this view, entrenchment would be operating at the level of how the constituent items are perceived or remembered, not upon the mechanism of statistical learning itself. This possibility is supported in principle by research demonstrating how perception is tuned based on experience. Developmental research on the perceptual tuning of native versus nonnative sounds in speech (e.g., Werker & Tees, 1984, 1999) and music (Hannon, Soley, & Levine, 2011; Lynch et al., 1990; Lynch & Eilers, 1992; Soley & Hannon, 2010) has found an enhanced ability to perceptually discriminate native sounds at the expense of nonnative sounds. Extending these findings to the present study, it is possible that listeners exhibited a warped perception (cf. Kuhl, 1991) of the basic interval in the novel tonal system (240 cents), perceiving this interval in terms of its nearest Western interval category - i.e., either an out-of-tune major second (200 cents) or an out-of-tune minor third (300 cents). Listeners in our study were able to differentiate adjacent tones in the present system from their nearest Western interval categories for both familiar and artificial instruments (see *Supplemental Material S2: Experiment Assessing the Perception of Intervals*). However, it is still possible that these novel intervals were encoded and/or stored in long-term memory less precisely (and more in line with Western interval categories), particularly when the constituent tones had the timbre of familiar instruments. Indeed, even if listeners are able to perceptually differentiate two items, these items can still be remembered more categorically in the time since encoding (e.g., Heald, Van Hedger, & Nusbaum, 2014; Olsson & Poom, 2005). The present results are aligned with this possibility, and further indicate that the extent to which these memory representations are categorical and align with Western musical intervals depends on instrument familiarity, suggesting that this process is contextually driven.

Second, it is possible that prior knowledge shapes processing of *sequential relationships,* operating at the level of statistical learning per

se. If the artificial timbres used in this experiment were simply not heard as music, while the sequences presented in familiar timbres were heard more musically, listeners would process the two types of sequences under different sets of rules or assumptions. The study instructions did not refer to the tonal sequences as "music" as we did not want to bias how participants approached the learning task. Nevertheless, hearing these novel tonal sequences being played by a piano or violin might have encouraged the participants in the Familiar Instrument Group to approach the paradigm as a music-learning task. Listeners in the Familiar Instrument Group thus may have had more difficulty abstracting the sequential statistics of the sequences, as they do not align with typical Western music. This explanation would be in line with previous item effects observed in linguistic statistical learning paradigms (e.g., Siegelman & Frost, 2015). This possibility is also conceptually similar to research in degraded speech understanding, in which sinewave speech is attended to and processed differently based on whether it is actually interpreted as speech (e.g., Márcio, Silva, & Bellini-leite, 2020; Möttönen et al., 2006). This possibility is further supported by findings showing that statistical learning might operate differently across different domains – including in situations where the abstract statistical structure is the same but is interpreted as belonging to different domains (Tompson, Kahn, Falk, Vettel, & Bassett, 2019). Future research could test this idea by manipulating the instructions to highlight the "musical" nature of the sequences and assess whether artificial instruments, given this framing, would exhibit learning patterns that are more akin to those observed for familiar instruments.

5. Conclusion

Overall, the present results provide clear evidence that prior knowledge can influence statistical learning in a non-linguistic domain. Our findings have implications for understanding domain generality versus domain specificity in statistical learning (e.g., Conway & Christiansen, 2006). Specifically, our results emphasize how the relationship of the to-be-learned items to previous knowledge structures can result in more domain-specific and idiosyncratic patterns of results. These results suggest that statistical learning might be best conceptualized as a process in which representations are *continually updated* based on both long-term accumulated knowledge and more immediate exposure to new patterns.

These findings therefore have implications for how researchers conceptualize statistical learning more broadly. Even in paradigms in which prior knowledge is assumed to play a minimal or nonexistent role, researchers must acknowledge that stimuli are multidimensional and that experience with one dimension can affect learning relevant to a second dimension (e.g., Garner & Felfoldy, 1970; Herrmann & Johnsrude, 2018). Inconsistencies in statistical learning findings across paradigms and tested items may reflect multiple statistical learning abilities (e.g., see Doeller & Burgess, 2008; Endress, 2019) that are inextricably tied to the domains upon which they draw. Although the present results cannot conclusively determine whether statistical learning is best conceptualized as a singular versus multifaceted ability, it is clear that further specifying the ways in which prior knowledge influences statistical learning across domains is essential in refining our theories of how statistical learning operates and underlies long-term learning in language, music, and beyond.

Data availability

All data associated with this manuscript are available on Open Science Framework (doi: 10.17605/OSF.IO/FZMQE).

Acknowledgements

This research was supported in part by a Natural Sciences and Engineering Research Council (NSERC) Discovery Grant (2019-05132) to Laura Batterink. Stephen Van Hedger is supported by a Canada First Research Excellence Fund (CFREF) BrainsCAN Postdoctoral Fellowship.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2021.104949.

References

Agus, T. R., Thorpe, S. J., & Pressnitzer, D. (2010). Rapid formation of robust auditory memories: Insights from noise. *Neuron, 66*(4), 610–618. https://doi.org/10.1016/j.neuron.2010.04.014

Arthur, W., & Day, D. V. (1994). Development of a short form for the raven advanced progressive matrices test. *In Educational and Psychological Measurement, 54*(2), 394–403. https://doi.org/10.1177/0013164494054002013

Attneave, F., & Olson, R. K. (1971). Pitch as a medium: A new approach to psychophysical scaling. *The American Journal of Psychology, 84*(2), 147–166.

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

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge University Press.

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

Batterink, L. J., Reber, P. J., Neville, H. J., & Paller, K. A. (2015). Implicit and explicit contributions to statistical learning. *Journal of Memory And, 83*, 62–78. https://doi.org/10.1016/j.jml.2015.04.004

Bigand, E., & Poulin-Charronnat, B. (2006). Are we "experienced listeners"? A review of the musical capacities that do not depend on formal musical training. *Cognition, 100*, 100–130. https://doi.org/10.1016/j.cognition.2005.11.007

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English/**r**/and/**l**/: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America, 101*(4), 2299–2310. https://doi.org/10.1121/1.418276

Brooks, L. R., & Vokey, J. R. (1991). Abstract analogies and abstracted grammars: Comments on Reber. *Journal of Experimental Psychology: General, 120*(3), 316–323.

Busey, T. A., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin and Review, 7*(1), 26–48.

Chua, E. F., Hannula, D. E., Ranganath, C., Chua, E. F., Hannula, D. E., Ranganath, C., … Ranganath, C. (2012). Distinguishing highly confident accurate and inaccurate memory: Insights about relevant and irrelevant influences on memory confidence. *Memory, 20*(1), 48–62. https://doi.org/10.1080/09658211.2011.633919

Conway, C. M., & Christiansen, M. H. (2006). Statistical learning within and between modalities pitting abstract against stimulus-specific representations. *Psychological Science, 17*(10), 905–912. https://doi.org/10.1111/j.1467-9280.2006.01801.x

Creel, S. C., Newport, E. L., & Aslin, R. N. (2004). Distant melodies: Statistical learning of nonadjacent dependencies in tone sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*(5), 1119–1130. https://doi.org/10.1037/0278-7393.30.5.1119

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

Doeller, C. F., & Burgess, N. (2008). Distinct error-correcting and incidental learning of location relative to landmarks and boundaries. *Proceedings of the National Academy of Sciences of the United States of America, 105*(15), 5909–5914. https://doi.org/10.1073/pnas.0711433105

Durrant, S. J., Cairney, S. A., & Lewis, P. A. (2013). Overnight consolidation aids the transfer of statistical knowledge from the medial temporal lobe to the striatum. *Cerebal Cortex, 23*, 2467–2478. https://doi.org/10.1093/cercor/bhs244

Durrant, S. J., Taylor, C., Cairney, S., & Lewis, P. A. (2011). Sleep-dependent consolidation of statistical learning. *Neuropsychologia, 49*(5), 1322–1331. https://doi.org/10.1016/j.neuropsychologia.2011.02.015

Endress, A. D. (2019). Duplications and domain-generality. *Psychological Bulletin, 145*(12), 1154–1175. https://doi.org/10.1037/bul0000213

Endress, A. D., & Bonatti, L. L. (2016). Words, rules, and mechanisms of language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science, 7*(1), 19–35. https://doi.org/10.1002/wcs.1376

Endress, A. D., Dehaene-Lambertz, G., & Mehler, J. (2007). Perceptual constraints and the learnability of simple grammars. *Cognition, 105*, 577–614. https://doi.org/10.1016/j.cognition.2006.12.014

Erickson, L. C., Kaschak, M. P., Thiessen, E. D., & Berry, C. A. S. (2016). Individual differences in statistical learning: Conceptual and measurement issues. *Collabra, 2*(1), 1–17.

Ettlinger, M., Margulis, E. H., & Wong, P. C. M. (2011). Implicit memory in music and language. *Frontiers in Psychology, 2*(September), 1–10. https://doi.org/10.3389/fpsyg.2011.00211

Finn, A. S., & Hudson Kam, C. L. (2008). The curse of knowledge: First language knowledge impairs adult learners' use of novel statistics for word segmentation. *Cognition, 108*(2), 477–499. https://doi.org/10.1016/j.cognition.2008.04.002

Finn, A. S., & Hudson Kam, C. L. (2015). Why segmentation matters: Experience-driven segmentation errors impair "morpheme" learning. *Journal of Experimental*

Psychology: Learning, Memory, and Cognition, 41(5), 1560–1569. https://doi.org/10.1037/xlm0000114
Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science, 12*(6), 499–504.
Frost, R., Armstrong, B. C., Christiansen, M. H., & Armstrong, B. C. (2019). Statistical learning research: A critical review and possible new directions. *Psychological Bulletin, 145*(12), 1128–1153. https://doi.org/10.1037/bul0000210
Furl, N., Kumar, S., Alter, K., Durrant, S., Shawe-Taylor, J., & Grif, T. D. (2011). Neural prediction of higher-order auditory sequence statistics. *NeuroImage, 54*, 2267–2277. https://doi.org/10.1016/j.neuroimage.2010.10.038
Garner, W. R., & Felfoldy, G. L. (1970). Integrality of stimulus dimensions in various types of information processing. *Cognitive Psychology, 1*, 225–241.
Hannon, E. E., Soley, G., & Levine, R. S. (2011). Constraints on infants' musical rhythm perception: Effects of interval ratio complexity and enculturation. *Developmental Science, 4*, 865–872. https://doi.org/10.1111/j.1467-7687.2011.01036.x
Hannon, E. E., Soley, G., & Ullal, S. (2012). Familiarity overrides complexity in rhythm perception: A cross-cultural comparison of American and Turkish listeners. *Journal of Experimental Psychology: Human Perception and Performance, 38*(3), 543–548. https://doi.org/10.1037/a0027225
Hannon, E. E., & Trehub, S. E. (2005a). Metrical categories in infancy and adulthood. *Psychological Science, 16*(1), 48–55.
Hannon, E. E., & Trehub, S. E. (2005b). Tuning in to musical rhythms: Infants learn more readily than adults. *Proceedings of the National Academy of Sciences, 102*(35), 12639–12643.
Heald, S. L. M., Van Hedger, S. C., & Nusbaum, H. C. (2014). Auditory category knowledge in experts and novices. *Frontiers in Neuroscience, 8*(AUG). https://doi.org/10.3389/fnins.2014.00260
Herrmann, B., & Johnsrude, I. S. (2018). Attentional state modulates the effect of an irrelevant stimulus dimension on perception. *Journal of Experimental Psychology: Human Perception and Performance, 44*(1), 89–105.
Holmes, E., Domingo, Y., & Johnsrude, I. S. (2018). Familiar voices are more intelligible, even if they are not recognized as familiar. *Psychological Science, 29*(10), 1575–1583. https://doi.org/10.1177/0956797618779083
Hutchins, S., Roquet, C., & Peretz, I. (2012). The vocal generosity effect: How bad can your singing be? *Music Perception, 30*, 147–159.
Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., & Carlyon, R. P. (2013). Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychological Science, 24*(10), 1995–2004. https://doi.org/10.1177/0956797613482467
Kim, R., Seitz, A., Feenstra, H., & Shams, L. (2009). Testing assumptions of statistical learning: Is it long-term and implicit? *Neuroscience Letters, 461*, 145–149. https://doi.org/10.1016/j.neulet.2009.06.030
Kóbor, A., Janacsek, K., Takács, Á., & Nemeth, D. (2017). Statistical learning leads to persistent memory: Evidence for one-year consolidation. *Scientific Reports, 7*(760), 1–10. https://doi.org/10.1038/s41598-017-00807-3
Krumhansl, C. L. (1990). In *Cognitive foundations of musical pitch*. Oxford University Press.
Krumhansl, C. L., Bharucha, J., & Castellano, M. A. (1982). Key distance effects on perceived harmonic structure in music. *Perception & Psychophysics, 32*(2), 96–108. https://doi.org/10.3758/BF03204269
Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics, 50*(2), 93–107.
Kuhl, P. K. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences, 97*(22), 11850–11857.
Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience, 5*(November), 831–843. https://doi.org/10.1038/nrn1533
Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science, 255*(5044), 606–608.
Kuhn, G., & Dienes, Z. (2005). Implicit learning of nonlocal musical rules: Implicitly learning more than chunks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(6), 1417–1432. https://doi.org/10.1037/0278-7393.31.6.1417
Lahav, A., Boulanger, A., Schlaug, G., & Saltzman, E. (2005). The power of listening: Auditory-motor interactions in musical training. *Annals of the New York Academy of Sciences, 1060*, 189–194. https://doi.org/10.1196/annals.1360.042
Leung, Y., & Dean, R. T. (2018). Learning unfamiliar pitch intervals: A novel paradigm for demonstrating the learning of statistical associations between musical pitches. *PLoS One, 13*(8), Article e0203026. https://doi.org/10.1371/journal. pone.0203026
Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*(1), 1–36. https://doi.org/10.1016/0010-0277(85)90021-6
Lim, S. J., & Holt, L. L. (2011). Learning foreign sounds in an alien world: Videogame training improves non-native speech categorization. *Cognitive Science, 35*(7), 1390–1405. https://doi.org/10.1111/j.1551-6709.2011.01192.x
Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods, 49*, 433–442. https://doi.org/10.3758/s13428-016-0727-z
Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English/r/and/l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America, 94*(3), 1242–1255. https://doi.org/10.1121/1.408177
Lockhead, G., & 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
Loui, P. (2012). Learning and liking of melody and harmony: Further studies in artificial grammar learning. *Topics in Cognitive Science, 4*, 554–567. https://doi.org/10.1111/j.1756-8765.2012.01208.x
Loui, P., & Wessel, D. (2008). Learning and liking an artificial musical system: Effects of set size and repeated exposure. *Musicae Scientiae, 12*(2), 207–230.
Loui, P., Wessel, D. L., & Hudson Kam, C. L. (2010). Humans rapidly learn grammatical structure in a new musical scale. *Music Perception, 27*(5), 377–388.
Lynch, M. P., & Eilers, R. E. (1992). A study of perceptual development for musical tuning. *Perception & Psychophysics, 52*(6).
Lynch, M. P., Eilers, R. E., Oller, D. K., Urbano, R. C., Lynch, M. P., Eilers, R. E., … Urbano, R. C. (1990). Innateness, experience, and music perception. *Psychological Science, 1*(4), 272–276. https://doi.org/10.1111/j.1467-9280.1990.tb00213.x
Márcio, D., Silva, R., & Bellini-leite, S. C. (2020). Cross-modal correspondences in sine wave: Speech versus non-speech modes. *Attention, Perception, & Psychophysics*, 944–953.
Margulis, E. H., Mlsna, L. M., Uppunda, A. K., Parrish, T. B., & Wong, P. C. M. (2009). Selective neurophysiologic responses to music in instrumentalists with different listening biographies. *Human Brain Mapping, 30*(1), 267–275. https://doi.org/10.1002/hbm.20503
Mathews, M. V., Pierce, J. R., Reeves, A., & Roberts, L. A. (1988). Theoretical and experimental explorations of the Bohlen-Pierce scale. *The Journal of the Acoustical Society of America, 84*(4), 1214–1222. https://doi.org/10.1121/1.396622
Mcclelland, J. L., Mcnaughton, B. L., & Reilly, R. C. O. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102*(3), 419–457.
McMullin, E., & Saffran, J. R. (2004). Music and language: A developmental comparison. *Music Perception, 21*(3), 289–311.
Miyazaki, K. (1989). Absolute pitch identification: Effects of timbre and pitch region. *Music Perception: An Interdisciplinary Journal, 7*(1), 1–14. https://doi.org/10.2307/40285445
Möttönen, R., Calvert, G. A., Jääskeläinen, I. P., Matthews, P. M., Thesen, T., Tuomainen, J., & Sams, M. (2006). Perceiving identical sounds as speech or non-speech modulates activity in the left posterior superior temporal sulcus. *NeuroImage, 30*, 563–569. https://doi.org/10.1016/j.neuroimage.2005.10.002
Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics, 60*(3), 355–376. https://doi.org/10.3758/BF03206860
Olsson, H., & Poom, L. (2005). Visual memory needs categories. *Proceedings of the National Academy of Sciences, 102*(24), 8776–8780. https://doi.org/10.1073/pnas.0500810102
Patil, K., Pressnitzer, D., Shamma, S., & Elhilali, M. (2012). Music in our ears: The biological bases of musical timbre perception. *PLoS Computational Biology, 8*(11), 1–16. https://doi.org/10.1371/journal.pcbi.1002759
Core Team, R. (2016). R: A language and environment for statistical computing. *R Foundation for Statistical Computing. Vienna, Austria*. https://www.R-project.org/.
Rohrmeier, M., & Rebuschat, P. (2012). Implicit learning and acquisition of music. *Topics in Cognitive Science, 4*, 525–553. https://doi.org/10.1111/j.1756-8765.2012.01223.x
Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*(5294), 1926–1928.
Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition, 70*, 27–52.
Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology, 69*, 181–203. https://doi.org/10.1146/annurev-psych-122216- 011805
Schapiro, A. C., Turk-Browne, N. B., Norman, K. A., & Botvinick, M. M. (2016). Statistical learning of temporal community structure in the Hippocampus. *Hippocampus, 8*, 3–8. https://doi.org/10.1002/hipo.22523
Schellenberg, E. G., & Habashi, P. (2015). Remembering the melody and timbre, forgetting the key and tempo. *Memory and Cognition, 43*(7), 1021–1031. https://doi.org/10.3758/s13421-015-0519-1
Siegel, J. A., & Siegel, W. (1977). Categorical perception of tonal intervals: Musicians can't tell sharp from flat. *Perception & Psychophysics, 21*(5), 399–407.
Siegelman, N., Bogaerts, L., Elazar, A., Arciuli, J., & Frost, R. (2018). Linguistic entrenchment: Prior knowledge impacts statistical learning performance. *Cognition, 177*(August 2017), 198–213. https://doi.org/10.1016/j.cognition.2018.04.011
Siegelman, N., & Frost, R. (2015). Statistical learning as an individual ability: Theoretical perspectives and empirical evidence. *Journal of Memory and Language, 81*, 105–120. https://doi.org/10.1016/j.jml.2015.02.001
Soley, G., & Hannon, E. E. (2010). Infants prefer the musical meter of their own culture: A cross-cultural comparison. *Developmental Psychology, 46*(1), 286–292. https://doi.org/10.1037/a0017555
Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature, 388*, 381–382.
Strait, D. L., Chan, K., Ashley, R., & Kraus, N. (2012). Specialization among the specialized: Auditory brainstem function is tuned in to timbre. *Cortex, 48*(3), 360–362. https://doi.org/10.1016/j.cortex.2011.03.015
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
Thiessen, E. D. (2017). What's statistical about learning? Insights from modelling statistical learning as a set of memory processes. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 372*(1711). https://doi.org/10.1098/rstb.2016.0056
Tillmann, B., & Bigand, E. (2000). Implicit learning of tonality: A self-organizing approach. *Psychological Review, 107*(4), 885–913. https://doi.org/10.1037//0033-295X.107.4.885
Tompson, S. H., Kahn, A. E., Falk, E. B., Vettel, J. M., & Bassett, D. S. (2019). Individual differences in learning social and nonsocial network structures. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 45*(2), 253–271.

Tunney, R., & Altmann, G. T. M. (2001). Two modes of transfer in artificial grammar learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, May*. https://doi.org/10.1037/0278-7393.27.3.614

Turk-Browne, N. B., Junge, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General, 134*(4), 552–564. https://doi.org/10.1037/0096-3445.134.4.552

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 and Review, 24*(2). 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. *Journal of the Acoustical Society of America, 138*(1). https://doi.org/10.1121/1.4922952

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

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

Weiss, M. W., Trehub, S. E., & Schellenberg, E. G. (2012). Something in the way she sings: Enhanced memory for vocal melodies. *Psychological Science, 23*(10), 1074–1078. https://doi.org/10.1177/0956797612442552

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7*, 49–63.

Werker, J. F., & Tees, R. C. (1999). Influences on infant speech processing: Toward a new synthesis. *Annual Review of Psychology, 50*, 509–535.

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

Zhao, J., Ngo, N., McKendrick, R., & Turk-Browne, N. B. (2011). Mutual interference between statistical summary perception and statistical learning. *Psychological Science, 22*(9), 1212–1219. https://doi.org/10.1177/0956797611419304