Article

MUSICÆ SCIENTIÆ

Absolute memory for pitch: A comparative replication of Levitin's 1994 study in six European labs

Musicae Scientiae
17(3) 334–349

DOI: 10.1177/1029864913493802
msx.sagepub.com
SAGE

Klaus Frieler
Hochschule für Musik Franz Liszt Weimar, Germany

Timo Fischinger
Humboldt-Universität zu Berlin, Germany

Kathrin Schlemmer
Katholische Universität Eichstätt-Ingolstadt, Germany

Kai Lothwesen
Hochschule für Musik und Darstellende Kunst, Frankfurt/M., Germany

Kelly Jakubowski
Goldsmiths, University of London, UK

Daniel Müllensiefen
Goldsmiths, University of London, UK

Abstract

In a widely cited study, Levitin (1994) suggested the existence of absolute pitch memory for music in the general population beyond the rare trait of genuine absolute pitch (AP). In his sample, a significant proportion of non-AP possessors were able to reproduce absolute pitch levels when asked to sing very familiar pop songs from memory. Forty-four percent of participants sang the correct pitch on at least one of two trials, and 12% were correct on both trials. However, until now, no replication of this study has ever been published. The current paper presents the results of a large replication endeavour across six different labs in Germany and the UK. All labs used the same methodology, carefully replicating Levitin's original experiment. In each lab, between 40 and 50 participants were tested ($N = 277$). Participants were asked to sing two different pop songs of their choice. All sung productions were compared to the original songs. Twenty-five percent of the participants sang the exact pitch of at least one of the two chosen songs

Corresponding author:
Klaus Frieler, Department of Musicology, Hochschule für Musik Franz Liszt Weimar, Germany.
Email: klaus.frieler@hfm-weimar.de

and 4% hit the right pitches for both songs. Our results generally confirm the findings of Levitin (1994). However, the results differ considerably across laboratories, and the estimated overall effect using meta-analysis techniques was significantly smaller than Levitin's original result. This illustrates the variability of empirical findings derived from small sample sizes and corroborates the need for replication and meta-analytical studies in music psychology in general.

Keywords
absolute pitch, collaborative research, music listening, music memory, replication

Almost all humans have some degree of absolute pitch memory since they are able to classify a single tone as being "high" or "low" without further reference. In this sense, absolute pitch memory is a trivial fact, but the interesting questions are its resolution and precision. Classical absolute pitch (AP) definitions typically use a one semitone resolution, i.e., if you are able to label or to produce a pitch within 1 semitone of the correct pitch you are deemed an AP possessor. Absolute pitch has been estimated to exist in less than .01% of the general population (Bachem, 1955; Profita & Bidder, 1988) and in about 15% of professional musicians (Baharloo, Johnston, Service, Gitschier, & Freimer, 1998). AP has often been regarded as an "all-or-none" phenomenon, with a bimodal distribution in the human population (Athos et al., 2007), but this might be partly due to the very restrictive definition. Furthermore, there exists some evidence that a more latent form of AP may be widespread in the population (Halpern, 1989; Levitin, 1994; Schellenberg & Trehub, 2003; Schlemmer, 2009; Smith & Schmuckler, 2008; Terhardt & Seewann, 1983; Terhardt & Ward, 1982). In most of these studies, non-AP participants successfully retrieved the absolute pitches of familiar pieces of music from memory, a phenomenon that has been termed "residual AP" (Takeuchi & Hulse, 1993) or, more recently, "implicit AP" (Deutsch, 2013). The description of implicit AP is concurrent with a trend in the literature suggesting the widespread retention of other absolute features of music, including tempo (Levitin & Cook, 1996) and timbre (Schellenberg, Iverson, & McKinnon, 1999).

In an early study on implicit AP, Terhardt and Seewann (1983) tested 135 musicians on their memory for the tonality of the major key preludes of Bach's *Well-Tempered Clavier*. Participants heard the opening of one of the preludes in either the original key or a transposed version and were asked to judge whether the performance was in the original key, a lower key, or a higher key than the original. Non-AP musicians performed significantly above chance on this task, even when the transposition was only by one semitone, suggesting that they were able to make inferences about the performed key of a well-known piece based on AP information. The authors highlight the possibility for memory strategy differences between AP and non-AP possessors, speculating that "AP possessors primarily identify individual notes, while non-AP possessors unconsciously deduce from a series of notes a feeling of key" (Terhardt & Seewann, 1983, p. 63). This ability to deduce a feeling of key based on the overall pitch range of the piece, rather than labelling individual notes, has more recently been referred to as "global-relative pitch" (Creel & Tumlin, 2012).

The findings of Terhardt and Seewann (1983), which continued the work of Terhardt and Ward (1982), were generalised to a non-musician sample by Schellenberg and Trehub (2003). Forty-eight non-musician college students were asked to distinguish the original version of a familiar television theme song from a pitch-shifted version in a two-alternative forced choice paradigm. All pitch-shifted versions were 1 or 2 semitones above or below the original key. Participants performed significantly above chance level in both conditions, while performing significantly better at the 2-semitone shifted than the 1-semitone shifted condition. These

findings provide evidence for implicit AP within a non-musician sample, while corresponding to the general AP literature in which it is noted that AP possessors make 1-semitone pitch judgement errors more commonly than larger interval errors (Miyazaki, 1988). However, the difficulty of dissociating AP from global-relative pitch cues is still present in the task used in this study.

Smith and Schmuckler (2008) tested non-musician undergraduates on their ability to detect the normal telephone dial tone from pitch-shifted versions by classifying each stimulus as the "normal" dial tone, "higher than normal", or "lower than normal". Participants were able to perform this task significantly above chance level, with accuracy of judgements increasing as the degree of pitch shifting from the original increased. This suggests that the absolute pitch of an over-learned stimulus may be retained in memory even when the stimulus is not inherently musical in nature.

One commonality between the studies discussed thus far is that the paradigms employed relied on the passive recognition, rather than active retrieval, of pitch information stored in memory. These studies also do not clearly delineate between whether participants made their discriminations based on absolute pitch information or the overall, global pitch range of the stimuli used. In order to observe more precisely what information is actually retrieved from a participant's memory in a pitch memory task, a production task may be warranted.

Halpern (1989) studied implicit absolute pitch using a combination of recognition and production methods. In the latter, participants sang their preferred starting pitch of 8 familiar folk songs, such as "Twinkle, Twinkle Little Star" and "Yankee Doodle". Despite the fact that these songs are published and sung in many different keys, it was found that participants produced their starting pitches for each song at quite stable absolute pitch levels, both across the same session and between different sessions 48 hours apart. Halpern's findings suggest a fairly stable internal representation of the pitch of certain well-known melodies, even if they are not always heard in the same key. It is worth noting that Halpern (1989) tested the stability of individual participants' auditory imagery of a song, rather than actual pitch memory. The mental representations of these folk songs were not necessarily based on a particular recording or version of a song and a stable auditory image does not necessarily reflect an explicit memory of the song being heard in a particular key. Therefore, it is a different but very interesting question to investigate the accuracy of pitch memory for familiar songs that *have* typically been heard in only one key.

Perhaps the most widely cited study of implicit AP in the general population is Levitin's 1994 study of pitch production in non-musicians. In this study, 46 participants, all but two of whom self-reported as non-AP possessors, performed significantly above chance at reproducing the absolute pitches of well-known, self-selected pop songs. Forty percent of participants sang at the same pitch level as the original recording of a self-selected pop song on at least one of two trials. Twelve percent sang at the correct pitch level on both trials, and 44% sang within 2 semitones of the original recorded pitch on both trials. Based on these findings, Levitin proposed a two-component theory of absolute pitch ability, including pitch labelling and pitch memory. He suggested that "true" AP includes both components, while the pitch memory component – here referred to as implicit AP – has a greater incidence in the general population than previously recognised.

Following up on Levitin's (1994) results, Schlemmer (2009) asked musically educated participants to produce songs from memory, which they had previously sung in a choir or practised on their instrument. Thirty percent of the 40 instrumentalists sang a well-practised piece at the original pitch level, and 67% were within 1 semitone of the original pitch. Among the 138

choir singers, pitch memory was less accurate than the instrumentalists and depended on rehearsal intensity, while musical expertise and pitch labelling ability influenced accuracy in the production task for both singers and instrumentalists.

To date, Levitin's 1994 paper has been widely cited (99 citations in Web of Knowledge and 224 citations in Google Scholar as of May, 2013). The enthusiasm surrounding Levitin's findings stems in part from the fact that AP has been long believed to be an extremely rare and distinct ability. AP has been one of the core topics of interest in music psychology since its inception (e.g., Abraham, 1901; Meyer, 1899; von Kries, 1892) and the number of papers investigating AP and affirming the belief that it is a very rare ability is large (Deutsch, 2013). In this respect Levitin's (1994) paper has been highly influential in reversing, or at least modifying, some of the beliefs surrounding AP memory as an uncommon and unique skill. In addition, the paper describes the discovery of a robust skill in non-specialist participants. Thus, the results of the paper have implications for a large part of the general population, making "ordinary people" aware of a special auditory skill previously unknown to them. This positive message might have contributed to the impact of Levitin's paper beyond the scientific discourse.

However, despite its large impact and frequent citations, Levitin's (1994) study has never been thoroughly replicated. Levitin's findings are at odds with the large number of papers that report AP to be a very rare ability, and hence the currently assumed incidence rate of implicit AP is based on only a single sample (with $n = 46$) from 1994. The small data basis and the lack of thorough replication, which seem incommensurate with the large impact of the results reported by Levitin, have served as a starting point for the present study. Thus, the current study aims to replicate Levitin's original procedures and methods as closely as possible with a new sample population across six different laboratories. Our replication study aims to achieve four specific goals. First, we wish to establish whether implicit AP for well-known songs (still) exists in the general population. Second, we aim to provide a broader empirical basis from which a more precise and reliable incidence rate of implicit AP in the non-specialist population can be estimated. Third, we aim to assess Levitin's claims regarding the independence of implicit AP from gender, age, musical training, and other measures of musical engagement (no significant effects of any of these factors were found in the original study). Finally, we aim to set an example for replication studies in music psychology. As Frieler et al. (2013) point out, replication studies on effects and phenomena in music psychology are very rare and in order for music psychology to establish itself as a serious scientific discipline that produces robust results, a culture of empirical replication needs to be introduced.

Method

The replication study was carried out in six different European labs, namely in the music or musicology departments of the Universities of Hamburg, Kassel, Eichstätt, and Frankfurt, and in the psychology departments of the Humboldt University, Berlin, and Goldsmiths, University of London. All labs used the same methodology, carefully replicating the experimental conditions of Levitin's (1994) study.

Participants

Altogether, 295 participants were tested, but only a subset of 277 (162 females, 115 males) participants delivered at least one usable trial; 250 subjects delivered two usable trials. The number of participants with at least one usable trial in the six labs was as follows: Hamburg, *n*

= 46, Kassel, n = 46, Eichstätt, n = 47, Frankfurt, n = 44, Berlin, n = 50, and London, n = 44. Participants were mostly students, unselected for musical background. The mean age of the participants was 25.85 (SD = 8.56) years. When asked to describe their musical background, participants classified themselves as follows: 10 non-musicians, 99 music-loving non-musicians, 50 amateur musicians, 51 serious amateurs, 51 semi-professional musicians, and 16 professional musicians (categories taken from Ollen, 2006). Seven participants claimed to possess AP. Table S1 summarises data on participants' musical backgrounds (see supplementary online section).

Materials

In order to realise the same experimental conditions in each lab, an online questionnaire (soscisurvey.de) was used that included all instructions for the entire experiment (see questionnaire in the supplementary online section). This online questionnaire was read aloud to the participants by the experimenter, who filled in the answers according to the information provided by the participants. The instructions for the production task closely resembled Levitin's original instructions, which were translated into German for the 5 German laboratories. The questionnaire included all questions used by Levitin and a few additional questions regarding the produced songs, for instance, if participants remembered when they had last heard these songs or if they had ever performed one of these songs (e.g., in a band). Additionally, the short form of Ollen's Musical Sophistication Index (Ollen, 2006) was used in order to gather standardised data on participants' musical backgrounds.

With respect to the songs from which participants could choose, one aspect of the current procedure differed from Levitin's study. Levitin had presented 58 compact discs (CDs) containing altogether more than 600 songs (from 58 artists and bands) with the intention of providing participants a "visual cue for subsequent auditory imaging" (Levitin, 1994, p. 416). Since in the 21st century, music is not primarily listened to on CD, but through various media including data files on the computer (such as mp3s), no CDs were presented in the experiment. Instead, participants were allowed to produce any song that they were very familiar with, or to choose a song from a provided song list. The song list contained 145 songs from 86 different artists and bands, which were compiled from lists of the most popular songs of the last 50 years as well as of best-selling songs per year from the last 20 years.

Following Levitin's original instructions participants were asked to first imagine the tune they wanted to produce. Participants' productions were recorded with a digital recording device and transferred to the computer as wav-files for analysis. Participants' answers on the questionnaire were recorded online and saved to a common spreadsheet at the end of the individual data collection phases.

Procedure

Participants were tested individually. At the beginning of the experiment, they were instructed to recall a song from memory that they knew very well, and to try to hear this song "playing in their heads". Then they were asked to sing, hum, or whistle a self-selected part of the song. If a participant could not think of an appropriate song, they were given the song list from which they could choose a song. Participants were allowed to repeat the production if they were unsatisfied with their performance. They were given no feedback and did not listen to their recorded productions. Participants then answered questions about the chosen song and filled

out the musical background questionnaire before the initial procedure was subsequently repeated for the second song trial. All participants were tested either in a quiet lab or, in 109 cases, in a quiet room at home from which all telephones had been removed. An experimental session lasted approximately 20 minutes.

Data analysis

The song productions were analysed four times. The first analysis was carried out manually, comparing the produced pitches to a precisely tuned musical instrument (e.g., digital piano, guitar, tuning fork), and assisted by the program Sonic Visualiser (Cannam, Landone, & Sandler, 2010). The second analysis was carried out by means of a pitch-tracking algorithm (De Cheveigné & Kawahara, 2002; Mauch, 2012), which determined the frequencies and pitch names of all produced pitches. In both analyses, the first well recognisable pitch produced by the participant was identified and rounded to the nearest semitone (in reference to a 440Hz tuning) as performed in Levitin (1994). A third manual analysis was conducted to resolve any discrepancies between the manual and automatic analyses. In the fourth step, an independent rater (not part of any of the research teams) re-analysed all productions in which the first two analyses had yielded different results, again manually and with Sonic Visualiser. Additionally, the independent rater analysed 10 randomly chosen productions per lab. All productions for which the analyses did not result in the same pitch category for the starting pitch were subsequently excluded from further analyses.

Beforehand, we had excluded all productions from analysis where the quality of the recording or the performance was not sufficiently high enough to recognise the target melody, or where the sung/hummed/whistled pitches were too unstable to reliably extract pitch values. In cases of doubt, trials were rigorously excluded to achieve very high data quality. Table S2 gives an overview of the number of trials excluded per lab.

In order to identify the original recorded pitches of the songs selected by participants, the original version of each song was acquired via Spotify, iTunes, etc. (as well as in some cases the original score). The pitches of the original song that corresponded to the phrase a participant had produced were analysed manually, assisted by Sonic Visualiser, analogously to the manual analysis of the participants' productions. All original recordings were analysed once in one of the labs and subsequently by the same independent rater who had cross-checked the participants' sung productions. The pitch difference between the participant's first sung pitch and the corresponding pitch from the original recorded version was measured in semitones and then served as the dependent variable for the statistical analyses. This analysis of the production data is comparable to Levitin's analysis in every respect, since both an automatic and a manual pitch identification procedure for identifying the starting pitches were carried out.[1]

Results

In Levitin's analysis, the pitch differences were mapped to pitch class differences (i.e., octave errors were not penalised), to compensate for possible octave transpositions used by the singers. In contrast to what is described in Levitin's paper, we used a numerical presentation of -6 to +5 to represent pitch differences in semitones, whereas he used a -6 to +6 range, splitting cases arbitrarily in the two end categories, which are mathematically identical. Because of the circular nature of the pitch class difference distributions we used circular statistics (Jammalamadaka

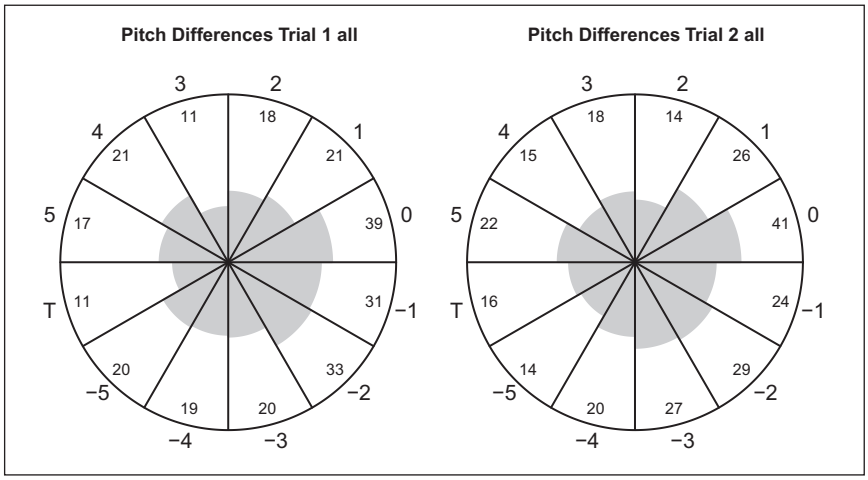

Figure 1. Wrapped pitch differences in semitone deviations between all subjects' productions and corresponding original pitches. Trial 1 (*M* = -1.00 semitones, *SD* = 3.39); Trial 2 (*M* = -.91 semitones, *SD* = 3.57).

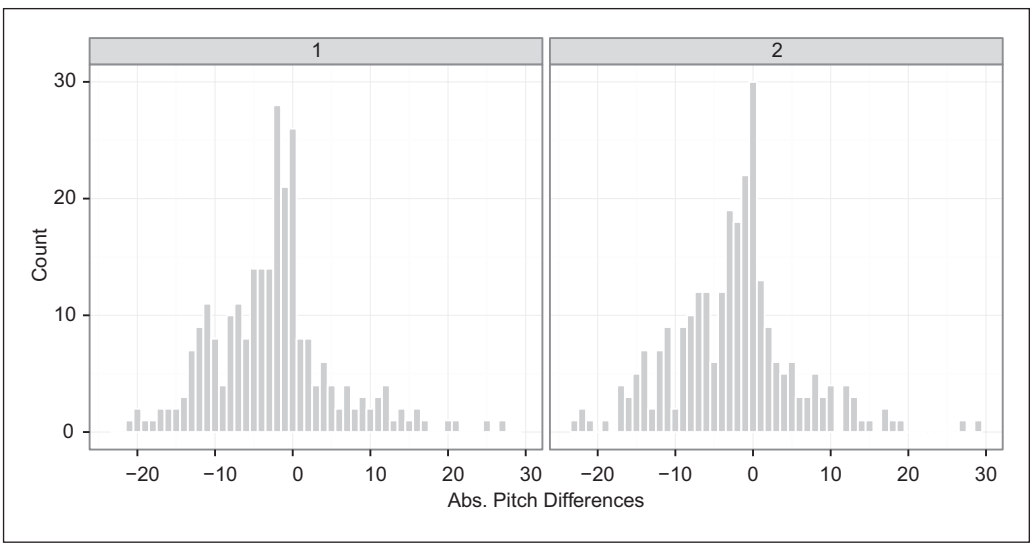

Figure 2. Absolute pitch differences in semitone deviations between subjects' productions and corresponding original pitches. Trial 1 (*M* = -2.65 semitones, *SD* = 7.92); Trial 2 (*M* = -2.43 semitones, *SD* = 7.92).

& SenGupta, 2001) to compute descriptive and inferential analyses in accordance with Levitin's original analyses. The circular mean of the pitch differences was -1.00 semitones (*SD* = 3.40) on Trial 1, and -.91 semitones (*SD* = 3.57) on Trial 2. The distributions of pitch class differences are displayed in Figure 1 (Trial 1 and Trial 2), those for absolute pitch differences in Figure 2. To test whether the pitch differences differed significantly from a uniform distribution we used Rayleigh's test which indicated significant differences for both trials (Trial 1: $r = .21$, $p < .001$, $N = 266$; Trial 2: $r = .17$, $p = < .001$, $N = 261$). For more detailed information about the

Table 1. Contingency table of "Hits" (= zero semitone difference) and "Misses" (all other differences) for Trials 1 and 2.

		Trial 2		Row Total
		Hits	Misses	
Trial 1	Hits	11	27	38
	Misses	30	182	212
	Column Total	41	209	250

Note. Only 250 subjects who delivered two usable trials were included.

Table 2. Contingency table of probabilities for "Hits" (= zero semitone difference) and "Misses" (all other differences) for the present (top) and Levitin's study (below).

	Present study	Trial 2	
		Hits	Misses
Trial 1	Hits	.04	.11
	Misses	.12	.73
	Levitin (1994)	Trial 2	
Trial 1	Hits	.12	.16
	Misses	.12	.60

statistical analyses for all labs and trials, see Table S3 in the supplementary online section. For further analysis, we defined trials with zero semitone pitch differences as "hits", whereas "close hits" were trials in which participants sang within +/- 1 semitone of the original pitch and "far hits" were productions within +/- 2 semitones of the original pitch.

On Trial 1, 39 of 261 subjects (15%) produced hits; 91 subjects (35%) close hits, and 142 subjects (54%) far hits. On Trial 2, 41 of 266 subjects (15%) hit the original pitch; 91 subjects (43%) had close hits, and 134 subjects (50%) far hits.

Consistency across trials was measured by creating a 2 × 2 contingency table of hits and misses (separately for each of the three definitions of "hits") for all 250 subjects who completed both trials. The raw counts are displayed in Table 1. The association is significant as measured by Yule's Q ($Q = .42$, $p = .012$), i.e., those who scored a hit in the first trial were also likely to score a hit in the second trial. As can be seen from Table 1, 11 subjects hit the correct pitch on both trials (4%), while 68 subjects (27%) were able to hit the correct pitch on at least one trial. For close hits, 77 subjects (31%) achieved a hit on at least one trial and far hits were achieved by 192 subjects (77%) on at least one trial.

For a direct comparison of both the present data and Levitin's results, Table 2 illustrates the probabilities of hits and misses shown in two 2 × 2 contingency tables.

To determine the variability of results between the six labs, we conducted a series of meta-analyses on the proportions of hits versus misses. In Figures 3 and 4 the hit rates in Trials 1 and 2 respectively are displayed for each lab including the data from Levitin's study. Fixed and random effects model give essentially the same results, and we will refer to the random effects model in the following as being the more conservative. The overall estimated hit rate in both trials is .17 (Trial 1 95% CI: [.12, .21]; Trial 2 95% CI: [.12, .21]). Note that the baseline probability of scoring a hit is $p_0 = 1/12 = .083$ (see Appendix S1), hence the

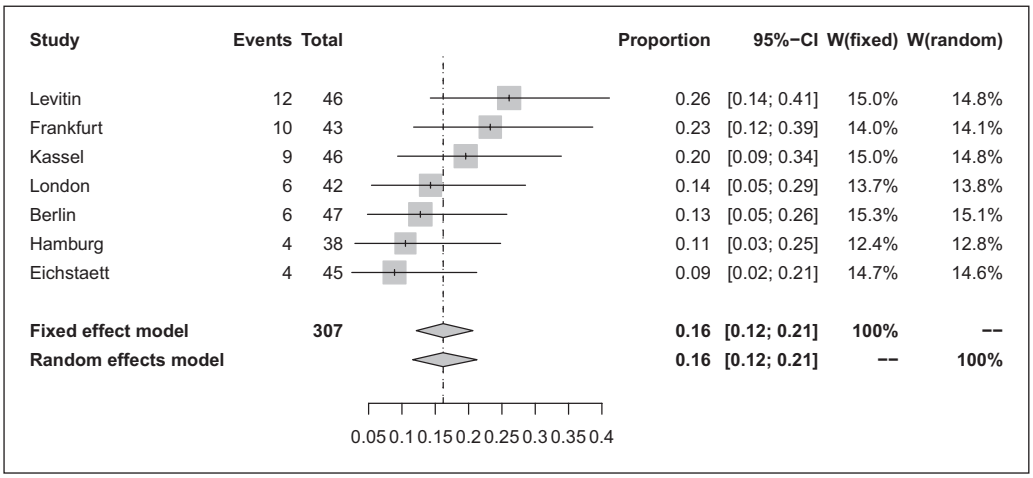

Figure 3. Meta analysis of "Hits" (= zero semitone difference between subject and original) for Trial 1 (Heterogeneity: I^2 = 23%, τ^2 = .0068, p = .2518).

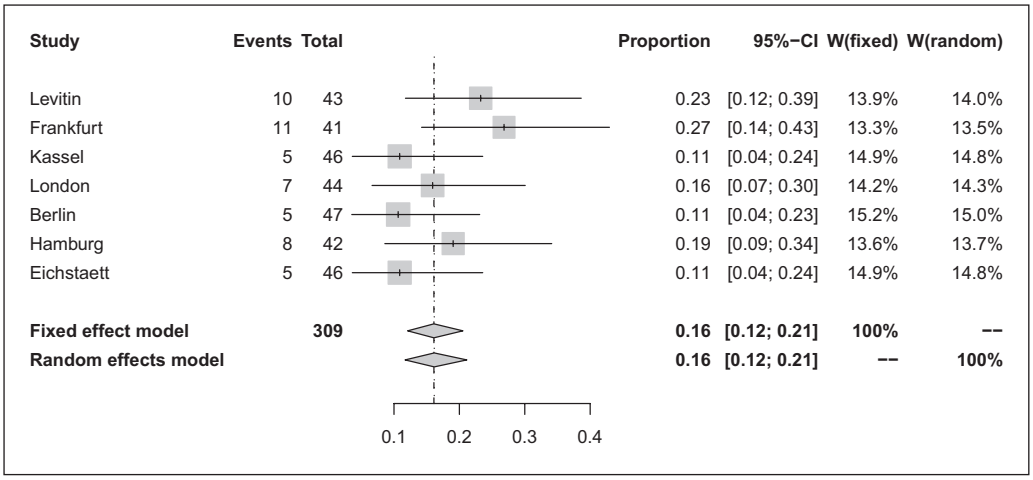

Figure 4. Meta analysis of "Hits" (= zero semitone difference between produced and original pitch) for Trial 2 (Heterogeneity: I^2 = 21%, τ^2 = .0058, p = .27).

measured overall hit rates are 2.04 times higher than the baseline. The heterogeneity of the different experiments is expressed by I^2, which amounts to 23.2% (p = .25) for Trial 1 and 18.9% (p = .26) for Trial 2, indicating that all experiments can be considered as measuring the same true effect. However, in both trials Levitin's original results are higher than in all other labs except Frankfurt, which attained even better results than Levitin in Trial 2. The Eichstätt lab is located at the other end of the spectrum, yielding results quite close to chance level. A third meta-analysis was conducted using "double hits" (hits on both trials) as the target variable (see Figure 5). Here the results are much more heterogeneous (I^2 = 62.7%, p = .013) with Levitin's and Frankfurt's samples performing much better than any of the

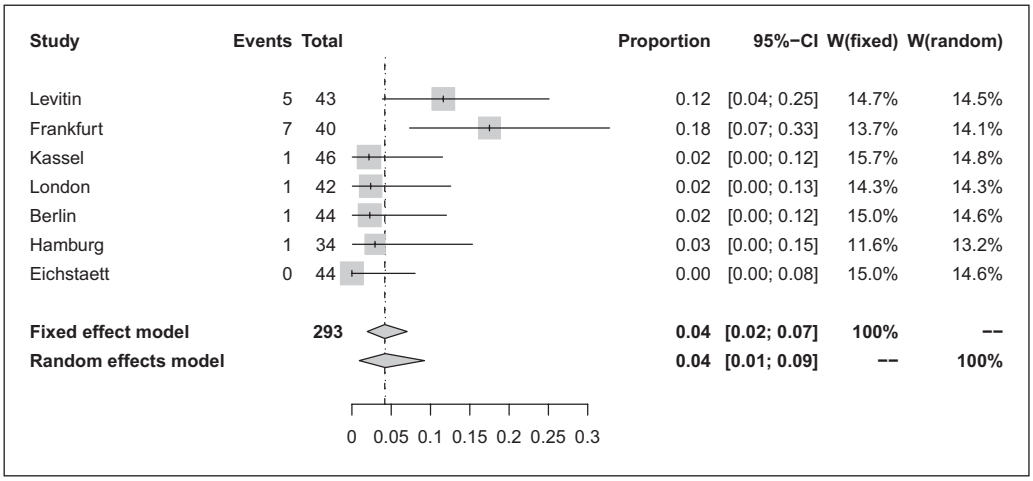

Figure 5. Meta analysis of "Double Hits" (= difference of zero semitones between produced and original pitch) on both Trial 1 and 2 for all different labs and Levitin's study separately (Heterogeneity: *I-squared* = 63%, *tau-squared* = .0393, p = .013).

other five labs. Note that the baseline probability of scoring a double hit is only $p_{00} = (1/12)^2 = .007$. In comparison, 12% of Levitin's participants scored double hits, which is 17.3 times higher than expected. Participants in the Frankfurt sample did even better with 18% double hits (25.9 times higher). All other labs show much lower double hit rates, but still higher than expected by chance. The overall estimated rate is .04, which is still 7.2 times higher than the baseline. Clearly, all meta-analyses demonstrate much higher overall estimated results than could be expected to be produced by chance alone.

To estimate power and effect sizes, we used a Goodness-of-Fit-χ^2-test to test the number of hits in both trials against the baseline probabilities. Using a significance level of $\alpha = .05$, Levitin's original study achieved an effect size of $\omega = .50$ and a power of .85 indicating a fairly strong effect and sufficient power. For our combined data, the statistical power was .98 and the effect size was only $\omega = .27$, which is roughly half of Levitin's effect size and only a medium effect. For a complete set of power and effect size estimates see Table S6 in the supplementary material.

Furthermore, we calculated correlations between hit rates on Levitin's production task with personal background variables collected in the accompanying online questionnaire. To this end, we used the total number of scored hits in both trials as the dependent variable. The procedure was as follows. For each categorical background variable a χ^2-Test was conducted (see Table S4) employing the False Discovery Rate algorithm for multiple testing (Benjamini & Hochberg, 1995). No variable showed any significant correlation after this correction, and only University Music Education (see Table S4 for a precise definition) ($\chi^2 = 20.91$, $df = 6$, $p = .002$, $p_{corr} = .091$) and explicit absolute pitch ability ($\chi^2 = 11.16$, $df = 2$, $p = .004$, $p_{corr} = .091$) remained weakly significant. Subsequently, each factor level of each variable was submitted to a χ^2-test against the baseline probabilities of number of hits ($p_0 = 121/144$, $p_1 = 22/144$, $p_2 = 1/144$). The results of this analysis are displayed in Table S5. The variables that showed the highest associations with hits in the production task were selected (and dichotomised where necessary) for three logistic regression models (mixed linear models including random effects), while also including demographic variables as independent variables, such as age, gender and

Table 3. Top 23 songs chosen for production in the present study.

Rank	Song	Artist	Count
1	Lemon Tree	Fools Garden	31
2	Let It Be	Beatles	22
3	Last Christmas	Wham!	16
4	Wonderwall	Oasis	16
5	Die perfekte Welle	Juli	15
6	Waterloo	ABBA	15
7	99 Luftballons	Nena	12
8	Tears in Heaven	Eric Clapton	9
9	My Heart Will Go On	Céline Dion	8
10	Someone Like You	Adele	8
11	Every Breath You Take	The Police	7
12	Piano Man	Billy Joel	7
13	All That She Wants	Ace of Base	6
14	Baby One More Time	Britney Spears	6
15	Feel	Robbie Williams	6
16	Imagine	John Lennon	6
17	Mrs Robinson	Simon & Garfunkel	6
18	Octopus's Garden	The Beatles	6
19	I Kissed a Girl	Katy Perry	5
20	Ironic	Alanis Morissette	5
21	Lady Madonna	The Beatles	5
22	Stairway to Heaven	Led Zeppelin	5
23	Viva La Vida	Coldplay	5

years of music lessons. The binary coding of hits, close hits and far hits on the basis of individual trials served as the dependent variable in each of the three regression models (see full data in the supplementary online section: Tables S7, S8 and S9). Subsequently, refined models using only significant terms from the first stage were calculated, but none of the included variables obtained statistical significance in the final models. All in all, this indicates that (close, far) hit rates are not a simple function of personal background variables, which in turn suggests that implicit AP can be seen as a rather general and complex phenomenon. This is in line with Levitin's results, which also did not contain any significant effects of gender, handedness, age, musical training, amount of time spent listening to music, or amount of time spent singing out loud. Nevertheless, our data hint at the possibility that musical sophistication and expertise may have some influence on hit rates. Participants with no university music education had a double hit probability of 3%, whereas musically educated participants had a rate of 9%. Still, both groups had much higher probabilities than baseline chance level (3.9 and 13.5 times higher, respectively), but after adjusting for multiple comparisons these differences were not statistically significant.

A final set of analyses investigated the relationship between the pitches sung on the two trials as well as the correct target pitches of the songs chosen for the two trials (see Table 3 for a list of the top 23 chosen songs). A highly significant correlation was found between the absolute semitone distances of the sung productions from the original recorded pitches across the two trials (Spearman's rho = .46, $p < .001$). Participants showed a tendency to produce the same absolute pitch difference from the original pitch in both trials (see Figure 6). In addition

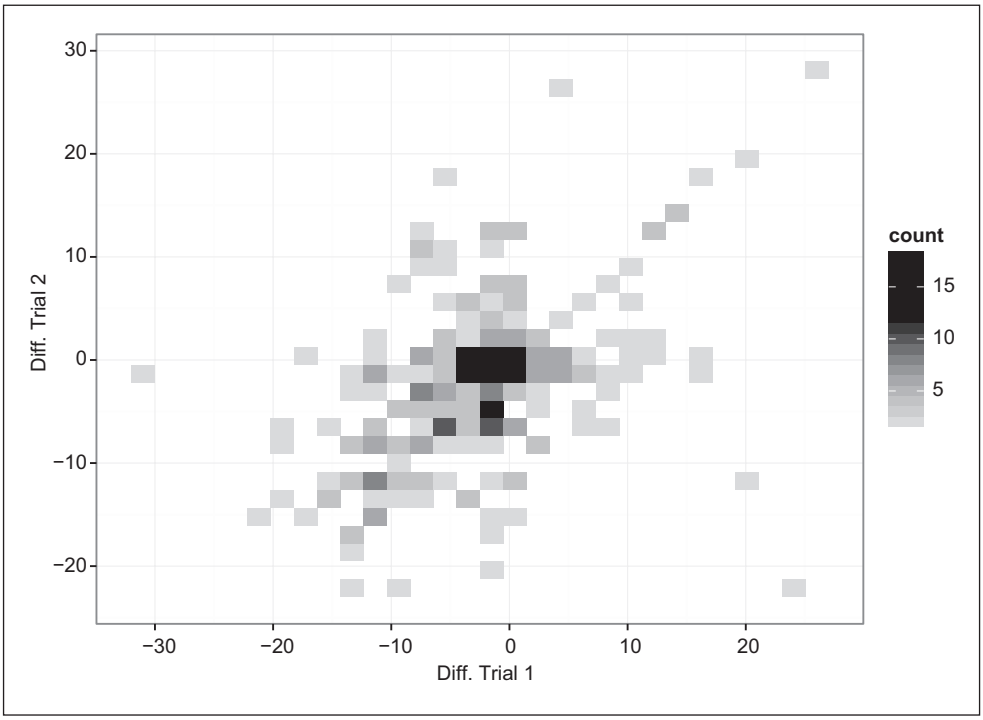

Figure 6. 2D-Scatterplot of absolute pitch differences in Trial 1 and Trial 2. Correlation is Spearman's rho = .46, $p < .001$.

we found an effect of the distribution of the target pitches in the original recordings that participants chose to sing. We mapped the target pitches onto pitch classes (0–11, with C mapped to 0) and conducted a Rayleigh test. The distribution of correct pitches differs very significantly from circular uniformity ($r = .19$, $p < .001$). The histogram (Figure 7) shows a preference for the pitch class C, as well as the neighbouring semitones and the tones of the C major triad. We have no fully satisfying explanation for this phenomenon. However, it might be partly explained by the fact that the distribution of pitch classes derived from 2855 melodies randomly sampled from a corpus of 14,000 commercially successful western pop songs (Müllensiefen, Wiggins, & Lewis, 2008) shows a preference for pitches of the C major pentatonic scale and that the distribution is significantly different from the distribution of target pitches ($\chi^2(99) = 108$, $p = .252$). Furthermore, a relationship between songs chosen on Trial 1 and Trial 2 was indicated by a significant correlation of the absolute pitch values of the target notes from both trials (Spearman's rho = .154, $p = .015$), and a significant correlation of pitches actually produced on both trials (Spearman's rho= .613, $p < .001$). We asked the subjects if they had previously performed the songs chosen for production (e.g., in a band), which allowed us to check whether performance experience had any influence on the hit probabilities. To this end, we conducted six χ^2-tests with the three logic variables hits, close hits and far hits vs. performed/not performed for Trial 1 and Trial 2, but none of the tests reached significance (Hits Trial 1: $\chi^2(1) = 2.8621$, $p = .09$; Close Hits 1: $\chi^2(1) = 1.795$, $p = .18$; Far Hits 1: $\chi^2(1) = .0349$, $p = .85$; Hits Trial 2: $\chi^2(1) = .0158$, $p = .90$; Close Hits 2: $\chi^2(1) = 1.246$, $p = .26$; Far Hits 2: $\chi^2(1) = 2.203$, $p = .14$).

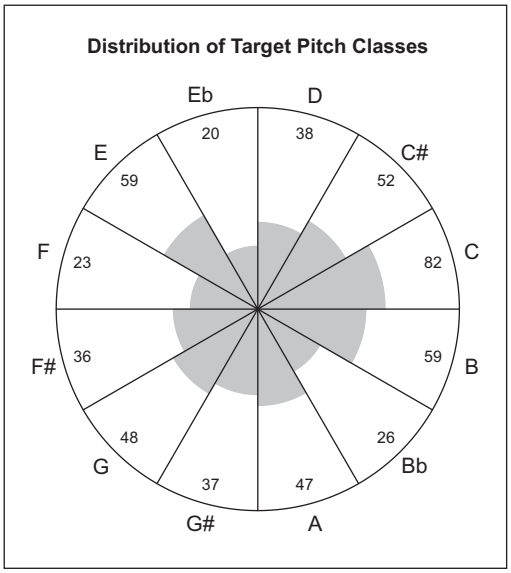

Figure 7. Histogram of pitch classes (0–11, with C mapped to 0) for the correct pitches in the study.

Finally, we analysed the performance of the 7 self-declared AP possessors in our study and found that only two of them scored double hits, whereas the remaining 5 did not score any hits in either of the trials.

Discussion

Given that 25% of the 277 participants who delivered at least one usable trial sang the exact pitch class on at least one song, and that 4% of the 250 who delivered two usable trials hit the right pitch of both songs, our results generally support the hypothesis of a latent absolute component for pitch memory for well-known tunes for a significant proportion of the general population, in accordance with Levitin (1994). However, our results differed across laboratories, and the overall estimated hit rates are much lower than in the original study (4% vs. 12% respectively). Despite the variation, all labs showed an effect in the predicted direction. The lower effect size might be caused by a) a decline effect (regression to the mean), b) changes in music listening behaviour over the last 20 years, which were partly a motivation for c) minor deviations from Levitin's method (see Method section). As for changes in music listening behaviour, we see one potentially important issue in the fact that many people (especially within the younger generation from which our sample is predominantly drawn) now listen to music via online streaming websites, such as YouTube. In doing this they might encounter different versions of the same song, differing – among other factors – in absolute pitch level. We aimed to control for this by asking each participant for the song version he or she was familiar with, but cannot rule out that even the same version (same singer) is in some cases available in different keys.

Additionally, the performance of the self-declared AP possessors in our study seems to indicate that even if the ability to label pitches is present, this does not necessarily imply that the pitch of a particular song is remembered correctly and that a correctly remembered pitch is actually used for reproduction. Particularly, the second point is important since the choice of a comfortable pitch might overrule the need for correct reproduction. If these considerations are

correct, it means that the actual share of people with implicit AP memory might be higher than estimated here, because the performance rates would have to be adjusted for false memories and the constraints of sung productions.

On the other hand, non-uniform and narrow distributions of preferred pitches in the participants as well as in the pop songs chosen could theoretically introduce a bias towards increased hit rates (imagine the difference between two dice for which the probability of a zero difference is higher than all other possible differences). To check for this, we conducted a set of Monte Carlo simulations with different possible estimates for the distribution of preferred and target pitches (see Appendix S1 and Table S10 for full details). The results of these simulations readily indicate that a bias of this kind seems very unlikely, hence, the baseline probability for a hit can safely be assumed to be the guessing probability of 1/12.

Follow-up experiments on implicit AP could employ singing tasks where comfortable pitch ranges are controlled for in order to mitigate individual singing deficiencies, pitch production paradigms other than singing, and perception tasks (e.g., adjusting a pitch via a dial). Recognition methods might implicate other disadvantages, e.g., it is very hard to control for familiarity and previous exposure to the stimuli and one cannot guarantee that experimenter-selected songs are encoded as well in memory as the individual participants' favourite songs. Rather, the experimenter must select a few songs that are presumed to be very familiar to the general population, which are then presented in different versions (see Schellenberg & Trehub, 2003; Terhardt & Seewann, 1983; Terhardt & Ward, 1982).

In addition, controlling the choice of songs for subsequent trials could help to clarify the correlations we found between the starting pitches of the two songs chosen and the pitches produced in both trials. The free choice of songs in our study might partly explain the observed double hit rates. The data seems to imply that at least some participants tend to choose and recall songs with a certain preferred pitch (cf. Moelants, Styns, & Leman, 2006), even though they may be unaware of it. We have no satisfying explanation for this rather puzzling result at this point, but it seems to constitute a systematic effect related to absolute pitch memory.

The present findings are not indicative of any particular conceptual model of implicit AP. There are two main theoretical options: (1) everyone has a general ability to remember absolute pitches correctly, or (2) implicit AP is actually like true AP – an all-or-none ability with a share of undetected implicit AP possessors who, as Levitin suggested, are able to memorise pitches but are not able to label them. From the present findings neither possibility can yet be ruled out.

Taken together, the results of this replication study indicate a considerable variability in the incidence rate of implicit AP. This variability is made explicit by the large confidence interval (ranging from 2% to 10%) for the estimate of the "true" proportion of the population that is able to sing correctly on both trials when all the data, including the data from Levitin's study, is taken into account. The data from five out of six labs reported here showed much smaller proportions compared to the original study. The largest discrepancy was found for the double hit rates, where the proportions of participants with double hits from the Frankfurt and Levitin samples lie outside even the confidence interval of the true population proportion. Because the Frankfurt sample was the only one where participants were recruited from a music university and because we found some association between task performance and higher music education we suspect a causal link here. However, this link does not apply to Levitin's 1994 results, where the sample included predominately students from disciplines other than music. Thus, while in this replication study we have been able to confirm that implicit AP is a much more frequent ability compared to genuine AP pitch labelling, the correlates and causes of implicit AP still remain to be discovered (see Jakubowski & Müllensiefen, 2013, for an extension of the present project investigating some additional correlates of implicit AP).

Conclusions

Our experiments support the claim that a general phenomenon of implicit AP is existent. But the comparison of the results of the six labs and Levitin's results show that replication of important and seminal experiments is necessary to obtain more valid and reliable estimates. Even though Levitin's study can be considered as having sufficiently high power, the effect size dropped considerably (about 50%) with our 6 times larger sample size. Hence, a substantial replication and meta-analyses as presented here increase the overall reliability and validity and provides more accurate estimates. This is a line of action we would like to strongly encourage for the whole field of music psychology, where many important studies eagerly await thorough replication.

Acknowledgements

We thank Daniel Levitin for providing his original questionnaire, instructions, and title list, and for answering our questions on his experimental methods. We would also like to thank Katharina Bauer for cross-checking the raw pitch data. Last but not least, we would like to thank Hannes Brix, Franziska Jabs, Maximilian Pannenberg, Tobias Knickmann, Jan-Hendrik Müller, Philipp Reddig, Lukas Rowitz, Theresa Tamoszus, Katharina Kanthak, Peter Zanker, Peter Sedlmeyr, Dominik Greguletz, Benjamin Schmid, Alexander Rühl, Leonie Polster, Franziska Ostermeier, Fabio Dick, Annika Rüschoff, Susanne Traber, Johanna Geißel, Merle Johanna Tegtmeier, Florian Bruntz, Bastian Wagener, Kristin Hoffmann, Nils Graumann, Judith Eisel, Fabian Kinzl, and all other student assistants in our labs for their help in data collection and data analysis.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Supplemental material

Supplemental appendices, tables and figures are available on the journal website: msx.sagepub.com/supplemental.

Note

1. Additionally, we included a second method of analysis, in which for each production the first pitch of the first stable phrase was identified and its pitch category recorded. A stable phrase was defined as one in which at least 3/4 of the intervals were produced correctly. The obtained pitches were compared with the original versions of the songs as described above, resulting in a second set of pitch differences. The pitch differences obtained with this method did not differ significantly (t = -.73, df = 1004.9, p = .464) from the pitch differences obtained with Levitin's (1994) method of identifying the first sung pitches. Therefore, all subsequent analyses were carried out using the pitch differences obtained on the basis of the first pitch that the participants produced for each song, which is identical to Levitin's procedure and also provided more usable observations for our datasets.

References

Abraham, O. (1901). Das absolute Tonbewußtsein [Absolute tone consciousness]. *Sammelbände der Internationalen Musikgesellschaft*, *3*, 1–86.

Athos, E.A., Levinson, B., Kistler, A., Zemansky, J., Bostrom, A., Freimer, N., & Gitschier, J. (2007). Dichotomy and perceptual distortions in absolute pitch ability. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(37), 14795–14800.

Bachem, A. (1955). Absolute pitch. *Journal of the Acoustical Society of America*, *27*, 1180-1185.

Baharloo, S., Johnston, P. A., Service, S. K., Gitschier, J., & Freimer, N. B. (1998). Absolute pitch: An approach for identification of genetic and nongenetic components. *American Journal of Human Genetics*, *62*(2), 224–231.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57(1), 289–300.

Cannam, C., Landone, C., & Sandler, M. (2010). Sonic Visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the ACM Multimedia 2010 International Conference*, (pp. 1467–1468). Firenze, Italy: ACM. Retrieved from www.sonicvisualiser.org

Creel, S.C., & Tumlin, M.A. (2012). Online recognition of music is influenced by relative and absolute pitch information. *Cognitive Science*, *36*(2), 224–260.

De Cheveigné, A., & Kawahara, H. (2002). YIN: A fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, *111*(4), 1917–1930.

Deutsch, D. (2013). Absolute pitch. In D. Deutsch (Ed.), *The psychology of music* (3rd ed., pp. 141–182). San Diego, CA: Academic Press.

Frieler, K., Müllensiefen, D., Fischinger, T., Schlemmer, K., Jakubowski, K., & Lothwesen, K. (2013). Replication in music psychology. *Musicae Scientiae, 17*(3), 265–276.

Halpern, A.R. (1989). Memory for the absolute pitch of familiar songs. *Memory & Cognition*, *17*(5), 572–581.

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

Jammalamadaka, S.R., & SenGupta, A. (2001). *Topics in circular statistics, sections 3.3.2 and 3.4.1*. Singapore: World Scientific Press.

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

Levitin, D. J., & Cook, P. R. (1996). Memory for musical tempo: Additional evidence that auditory memory is absolute. *Perception & Psychophysics*, *58*(6), 927–935.

Mauch, M. (2012). *Tony: A practical note annotation software program*. Unreleased Matlab software. Retrieved from http://www.isophonics.net/tony

Meyer, M. (1899). Is the memory of absolute pitch capable of development by training? *Psychological Review*, *6*(5), 514–516.

Miyazaki, K. (1988). Musical pitch identification by absolute pitch possessors. *Perception & Psychophysics*, *44*(6), 501–512.

Moelants, D., Styns, F., & Leman, M. (2006). Pitch and tempo precision in the reproduction of familiar songs. Proceedings of the tenth International Conference on Music Perception and Cognition.

Müllensiefen, D., Wiggins, G., & Lewis, D. (2008). High-level feature descriptors and corpus-based musicology: Techniques for modelling music cognition. In A. Schneider (Ed.), Hamburger Jahrbuch für Musikwissenschaft, 24 (pp. 133–155). Frankfurt: Peter Lang.

Ollen, J.E. (2006). A criterion-related validity test of selected indicators of musical sophistication using expert ratings [Electronic resource]. Retrieved from http://www.ohiolink.edu/etd/view.cgi?osu1161705351

Profita, J., & Bidder, T.G. (1988). Perfect pitch. *American Journal of Medical Genetics*, *29*(4), 763–771.

Schellenberg, E.G., Iverson, P., & McKinnon, M.C. (1999). Name that tune: Identifying popular recordings from brief excerpts. *Psychonomic Bulletin & Review*, *6*(4), 641–646.

Schellenberg, E.G., & Trehub, S.E. (2003). Good pitch memory is widespread. *Psychological Science*, *14*(3), 262–266.

Schlemmer, K. (2009). Das Gedächtnis für Tonarten bei Nichtabsoluthörern: Einflüsse von Hörhäufigkeit und musikalischer Ausbildung [Memory for tonality in non-absolute-pitch-possessors: Influences of hearing frequency and musical education]. *Jahrbuch Musikpsychologie*, *20*, 123–140.

Smith, N.A., & Schmuckler, M.A. (2008). Dial A440 for absolute pitch: Absolute pitch memory by non-absolute pitch possessors. *Journal of the Acoustical Society of America*, *123*(4), EL77–84.

Takeuchi, A.H., & Hulse, S.H. (1993). Absolute pitch. *Psychological Bulletin*, *113*(2), 345-361.

Terhardt, E., & Seewann, M. (1983). Aural key identification and its relationship to absolute pitch. *Music Perception*, *1*(1), 63–83.

Terhardt, E., & Ward, W.D. (1982). Recognition of musical key: Exploratory study. *Journal of the Acoustical Society of America*, *72*(1), 26–33.

von Kries, J. (1892). Über das absolute Gehör [On absolute pitch]. *Zeitschrift für Psychologie*, *3*, 257–279.