Article:
Weir, G., Williamson, V. and Mullensiefen, D. (2015) Voluntary but not involuntary music mental activity is associated with more accurate musical imagery. Psychomusicology: Music, Mind and Brain, 25 (1). 48 - 57. ISSN 0275-3987

https://doi.org/10.1037/pmu0000076

Increased involuntary musical mental activity is not associated with more accurate voluntary musical imagery

Guy Weir

Department of Psychology, Goldsmiths, University of London

Victoria J. Williamson

Department Of Music, University Of Sheffield, United Kingdom

Hochschule Luzern – Musik, Lucerne University Of Applied Sciences And Arts, Switzerland

School of Advanced Study, University of London, United Kingdom

Daniel Müllensiefen

Department of Psychology, Goldsmiths, University of London

Daniel Müllensiefen (corresponding author)
Department of Psychology, Goldsmiths, University of London
New Cross Road, New Cross
London SE14 6NW
Telephone: +44-20-7919 7895
e-mail: d.mullensiefen@gold.ac.uk

Abstract

This study investigates whether there is an association between accurate performance on a musical imagery test and the extent to which people typically experience involuntary musical imagery (INMI, also known as 'earworms'). This hypothesis was tested alongside a second hypothesis regarding the established association between musical practice (musical training and activity) and musical imagery ability. Sixty-seven participants were recruited from a general adult population to represent groups with high/low everyday INMI experiences and high/low levels of musical practice. The experimental musical imagery task required participants to listen to excerpts of familiar songs that contained a muted section of about 10s and to judge whether the re-entry of the music after the muted section was shifted in pitch (up/down by 1 semitone) or in timing (early/late by 2 beats). Results confirmed the second hypothesis: musical practice was positively associated with the accuracy of pitch judgments on the imagery task but not timing judgments. By contrast, none of the INMI measures were associated with imagery accuracy. Results are interpreted with reference to the literature on expertise effects and musical imagery.

Keywords: Musical imagery; involuntary musical imagery; earworms; musical training; mental imagery

Introduction

'The soul never thinks without a mental image' Aristotle, c330 BC

Mental imagery is a prevalent psychological ability, and yet, it remains poorly understood compared to other forms of cognition. This situation is a consequence of the intractable nature of mental imagery, the historical unpopularity of introspective methods (Baddeley & Andrade, 2000), and the dominance of research on verbal processes for theory development (Kosslyn, 1994; Pylyshyn, 2002). The drive to understand mental imagery has, however, encouraged the development of creative experimental paradigms that measure aspects of mental imagery experience such as vividness, speed and accuracy (Shepard & Metzler, 1971; Funt, 1980; Baddeley & Andrade, 2000; Janata & Paroo, 2006; Lucas, Schubert & Halpern, 2010).

A typical voluntary mental imagery paradigm might require people to close their eyes and to imagine the sight of a kitten, the sound of a friend's voice, or the smell of freshly baked bread. The subsequent internal mental experience can be described as voluntary, since initiation and cessation of the imagery is under conscious control. The experience can also be compared to the corresponding perceptual occurrence, in the absence of any environmental stimulus.

Voluntary musical imagery ('VMI' hereafter) has been explored using similar paradigms. An example VMI trial might require a participant to state whether the fifth note of the song "Happy Birthday" was higher, lower or the same as the sixth note. A participant would have to use their 'mind's ear' to solve the VMI query, making no overt musical sound (Halpern, 1989, 1992; Smith, Reisberg & Wilson, 1992). Simple VMI judgments such as these

are favored over complex phenomenological reports of musical mental imagery experiences; they are more replicable, adaptable to neuroimaging, and theoretically informative (Bailes, 2007; Halpern, 1988a&b, 1989; Janata & Paroo, 2006; Reisberg, 1992; Webber & Brown, 1986; Zatorre & Halpern, 2005; Zatorre, Halpern, & Bouffard, 2010).

Trained musicians are known to employ VMI as part of their practice and performance habits, a process akin to notational audiation (Bailes, Bishop, Stevens & Dean, 2012; Brodsky et al., 2008). Aleman et al. (2000) established a link between an individual's level of musical expertise and their VMI ability, finding that musicians outperformed a matched group of musically naïve participants on a VMI pitch test (comparing pitches, as in the earlier example of "Happy Birthday") but not a similar visual imagery task. This result is in line with more recent findings reported by Bishop, Bailes and Dean (2013) using a similar experimental paradigm and demonstrating a clear association between the accuracy of loudness imagery for musical stimuli and musical expertise.

The experience of VMI in both musicians and non-musicians stands in stark contrast to the experience of Involuntary Musical Imagery ('earworms': INMI hereafter). This is a term given to spontaneous, unbidden and repeating excerpts of music in 'the mind's ear', a common mental phenomenon that occurs in the absence of direct volitional control (Beaman & Williams, 2010; Liikkanen, 2008: 2012; Halpern & Bartlett, 2011; Williamson et al., 2012; 2014).

The premise of the present study was to test whether the regularity of INMI experienced in daily life is associated with improved accuracy in VMI judgments. Motivation for this prediction is drawn from literature on auditory memory. Repetition, such as that experienced with musical imagery in extreme INMI frequency, is one mechanism by which information may be successfully stored in memory (Atkinson & Shiffrin, 1968; Baddeley &

Hitch, 1974). Voluntary repetition of musical imagery within memory is one reason why musicians may develop more accurate VMI pitch ability, though other explanations include greater attention focus at the point of encoding and more robust storage thanks to advanced knowledge of musical structure and notation (Bailes et al., 2013). To date, the question of whether involuntary repetition of musical material in the form of INMI also has a positive effect on VMI accuracy remains largely unexplored. Such a positive effect could suggest a link between the INMI phenomenon and involuntary repetition mechanisms, and could therefore contribute an answer to questions regarding the function and purpose of INMI.

Thus, the first aim of the present study was to replicate the established association between musical practice and VMI pitch ability. We decided to use a different imagery test than the one used by Aleman et al., (2000), one that makes use of realistic music stimuli. We also extended the scope of the experimental paradigm used by Aleman et al. (2000) to include a test of VMI timing as well as pitch ability. The second aim was to assess VMI pitch and timing abilities in individuals who report INMI that is frequent in occurrence and persistent in time ('extreme' INMI hereafter; Hyman et al., 2013; Liikkanen, 2012; Müllensiefen et al., 2014a; Williamson & Jilka, 2014). The research question was, is extreme INMI, like musical practice, associated with more accurate VMI ability?

The VMI paradigm

The present study tested VMI ability in a paradigm whereby recordings of familiar pop songs were muted for a short period and participants asked to imagine the music continuing in their 'mind's ear'. Our paradigm has its origins in an fMRI study by Kraemer et al. (2005) who intersected familiar music with short periods of silence (see also Bailes & Bigand (2004), for

a behavioral version of this paradigm). They found that silence within familiar music elicited greater activity in the auditory association areas of the brain than did silence in unfamiliar music. Self-reports from the participants suggested that they were continuing to experience imagery of the familiar tunes during the silences.

Our paradigm aimed to extend and validate the findings from the existing musical imagery research (Aleman et al., 2000) by using recordings of familiar pop songs as in the above studies. This extension increases the ecological validity of the VMI approach as a whole and makes the task more engaging for participants, thereby achieving better data quality.

Stimuli in the paradigm were designed to ensure that short-term memory for harmony or pitch would not lead participants to the correct answer (see Method section). Thus, similar to the paradigms used by Aleman et al. (2000) and Halpern (1989, 1992), the only strategy available other than guessing was for participants to rely on their VMI of the songs, as the musical stimulus was entirely absent during the muted period. After the muted period a continuation of the music was presented either at the correct (unaltered) or incorrect (altered) pitch level and/or point in time. The participants' task was to identify any pitch and timing manipulations.

Method

This study tested the VMI abilities of people who experience extreme INMI compared against those who do not. In addition we tested the contribution of musicianship to VMI ability. A quasi-experimental design was used to assign between-subject factors; participants were sampled to represent the four groups arising from the combination factor levels of high/low INMI experience and high/low musicianship.

Participants

Participants were selected according to their data on the earwormery survey database (Williamson et al., 2012). They were divided into one of four groups according to whether or not they experienced extreme INMI (both frequent and persistent), and whether or not they were practicing musicians.

The 'extreme' INMI group comprised participants who reported experiencing INMI more than once a day for more than 3 hours a day, or constantly. People in the low INMI group fulfilled neither of these criteria. Practicing musicians were defined as individuals who classified themselves at least as 'amateur musician' on a 6-point self-classification scheme (see *Materials* below), indicating that they had gone through a period of intense music making at some point during their lives. In addition, individuals in this group reported that they were currently playing their instrument or singing for at least one hour a month. Non-musicians were defined as individuals who fulfilled neither of these criteria. Invitations to take part in the study were sent to 400 people randomly selected from the earwormery database who fulfilled the criteria for one of the four groups. 83 individuals responded to the invitation (21% response rate) and 67 participants completed the full test. Table 1 gives the demographic information for the groups.

Table 1. *Size of participant groups, distribution of age and gender.*

	Musician and extreme INMI	Non-musician and extreme INMI	Musician and low INMI	Non-musician and low INMI
N (valid responses)	16	18	18	14
Female	10	7	15	7
Mean age (std dev)	39.1 (12.7)	44.6 (13.4)	41.1 (14.9)	39.1 (12.8)

Materials

Individual differences

In addition to the group assignment, participants completed questionnaires that allowed us to apply a more refined assignment of their experiences at the analysis stage. As a measure of musical activity, we took four items from the Musical Behavior Questionnaire representing the factors Musical Practice, Music Professionalism, Listening Engagement, and Singing (Müllensiefen et al., 2014a, p. 328). More specifically, these four items measured: 1) Amount of musical practice done during a period of sustained musical activity (6-point scale with categories 0 = “Never practiced an instrument”, 1=“About an hour per month”, 2=”About an hour per week”, 3=”About 15 minutes per day”, 4=”About an hour per day”, 5=”More than

two hours per day"); 2) Self-assessed musicianship (6-point scale with categories 1="Non-musician", 2="Music-loving non-musician", 3="Amateur musician", 4="Serious amateur musician", 5="Semi-professional musician", 6="Professional musician"); 3) Current amount of attentive listening (6-point scale with categories 1="Never", 2="Once a week", 3="A few times per week", 4="Nearly every day", 5="Once or twice a day", 6="Several times per day"); 4) Self-rated ability to carry a tune (4-point scale with categories 1="Not at all", 2="Not very well", 3="Fairly well", 4="Very well").

In order to assess the nature of INMI experiences we used two items from the Involuntary Musical Imagery Questionnaire (Müllensiefen et al., 2014a) that relate to the frequency of INMI experiences (5-point scale with categories 1="More than once a day", 2="Once a day", 3="At least once per week", 4="At least once per month", 5="Less than once per month") and the length of INMI episodes (5-point rating scale with categories 1="Less than 10 minutes", 2="Between 10 minutes and half an hour", 3="Between half an hour and two hours", 4="Between 1 hour and 3 hours", 5="More than 3 hours").

Musical Stimuli

The new VMI ability test was programmed into the online survey system 'Qualtrics'. This system can play mp3 format sound files automatically when a web page loads. It also allows the programmer to hide the media player from the participant so they are unable to manually stop, pause or fast forward playback. All pages were set so there was no 'back' function and after the music stimulus finished playing the pages automatically forwarded to the next one. The main advantages of using an online testing interface is that it enables the recruitment of a larger sample size and better response rate, and the access to a more diverse population (i.e. participants could be based all round the country and not be required

to visit the lab). The main disadvantage is surrendering control of standardized lab conditions.

Excerpts from nine pop songs were selected as stimulus items. These songs were chosen from the bestselling downloads by decade, as recorded by the Official Charts Company (http://www.officialcharts.com/). They covered each decade from the 1950s through to the 2010s. We chose these highly popular songs in order to maximize the likelihood that each participant, regardless of their age, would have a good chance of being highly familiar with the songs. Details of the songs are listed in Appendix Table 1: all were up-tempo apart from 'Wonderful World' and were presented in the exact version that charted.

The order of the song presentation was randomized across participants. In order to reduce the overall length of the experiment we extracted an excerpt from each song that included what we presumed to be the most memorable parts. Excerpts were cut in musically meaningful places and the average length of the song excerpts was 56.28 seconds (range 30.8s – 89.6s).

Song manipulations were carried out in Cubase (V.4 for Mac). All pitch shifting used the MPEX3 algorithm set to 'poly complex' quality. The pitch shift algorithm altered the entire content of the sound file, essentially resulting in a shift in key. Timing shifts were done by hand, erasing sections of the song for approximately 10 seconds depending on how the bars and beats of the songs presented themselves for cut points.

A 10 second silence was chosen to reduce the influence of sensory memory, which has an upper limit of about 5 seconds (Darwin, 1972). Cuts were always made in musically meaningful places, i.e. at phrase boundaries and were only done at beat locations in the song. The music dropped out suddenly, and did not fade out. Care was taken to ensure that the last

chord heard when the music muted was different from the chord presented to the participant when the music returned, so that pitch judgments could not rely on short-term memory for absolute pitch height or harmony. The returning music, altered or unaltered, maintained pitch and tempo properties until it faded out at the end of each trial.

Experimental Conditions

For each song excerpt, nine versions were prepared that represented nine different experimental conditions. There were three pitch modifications ("-1" = pitch shifted flat by one semitone / "0" = pitch unchanged; "1" = pitch shifted sharp by 1 semitone) and three timing modifications ("-1" = timing shifted to come in 2 beats early / "0" = timing unchanged; "1" = timing shifted to come in 2 beats late). The combination of these 3x3 modifications gave rise to nine experimental conditions, which were selected at random for the nine trials of each participant. Each excerpt contained a section of around 10 seconds where the music was muted (for exact details see Appendix Table 1). After this silent passage the music that returned was subject to one of these nine modifications of pitch and/or timing.

Procedure

The protocol was conducted over the Internet and participants were asked to find 20 minutes clear from distraction in a quiet space, and to wear headphones.

The experiment involved nine trials, one for each experimental condition. In each trial a participant listened to around a minute of a well-known pop song to refresh their memory of it and mitigate effects of recent exposure. They then listened again to the same section of the song, only this time the music was muted after around 10 to 40 seconds, for around 10

seconds (see details in *Materials* and Appendix Table 1). During this silent time participants were asked to continue to follow the music as if it were still playing in their 'mind's ear'.

After the 10 seconds of silence the music returned in one of the nine experimental conditions (see *Materials*). As soon as the music finished the participant had to rate how early or late the music entry was in comparison with their imagination by moving a virtual slider on the screen. They used a similar slider to indicate whether the pitch was lower, the same, or higher than they imagined. We collected categorical data, with any slider movement one way or the other coded as early or late, flat or sharp. When the slider remained neutral the response was coded as unaltered. Finally, participants were asked how familiar they were with the song in general (1= "extremely familiar"; 2= "quite familiar"; 3= "not very familiar"; 4= "heard now for the first time"). No time constraints were placed on any of the judgments.

The nine songs were each presented in random order across participants. Experimental conditions were also paired with songs at random for each participant using the Qualtrics Randomizer function. After the test was concluded, participants were shown a debrief screen which thanked them, summarized the aims of the study, and gave them a contact email address in order to find out their results.

Results

Pitch and timing responses were scored as correct/incorrect for each trial. The VMI paradigm works on the assumption that participants make use of their memory of the song when performing the imagery task during the silent gap. We tested this assumption by comparing task performances for the lowest vs the higher levels of song familiarity. We used a binomial mixed-effects model (similar to a within-participants logistic regression model)

with participant as the random effect variable, correct responses to both pitch and timing as dependent variable, and familiarity recoded as a binary variable ('never heard the song before' vs all other levels of familiarity) as independent variable. Familiarity was a significant predictor in the regression model (β = -0.9971, z = -2.265 p = .024) indicating a lower probability for giving a correct response when the song was completely unfamiliar. This result supports the assumption that the VMI ability task measures musical imagery ability that relies on song memory: participants who are not familiar with a given song should not be able to form a mental image leading to a significantly poorer performance on the task.

Following this analysis, participants' responses were only scored if they stated that they were either fairly or extremely familiar with the song. Respondents with fewer than 5 of the 9 trials completed were excluded. As a result, three participants were removed from the analysis of the aggregated test scores.

An initial analysis

Three dependent variables were derived from the data by computing the proportions of correct responses across trials for each participant with respect to their timing as well as their pitch responses: VMI *pitch score* (correct answers of pitch responses only), VMI *timing score* (correct answers of timing responses only), and VMI overall score (i.e. correct answers for both pitch and timing in the same trial). Scores on these three variables have a theoretical range from 0 to 1.

Participants were more accurate in their pitch responses (mean = 0.673, SD = 0.264) than in their timing responses (mean = 0.434, SD = 0.186). Because the overall score represented observations where both responses were correct, this variable yielded the lowest average score (mean = 0.306, SD = 0.220). Taken together, all three accuracy scores were significantly different from chance level according to a binomial test (p-values < .001).

Chance levels were 0.333 for the VMI pitch and the timing scores and 0.111 for VMI overall score.

In the main part of the analysis we used six predictor variables, four relating to musical activity (current amount of attentive listening, self-rated musicianship, self-rated singing ability and self-rated practice time) and two relating to INMI extent (INMI frequency and INMI length). These six predictor variables were entered in three multiple regression models, one for each dependent variable. Only the model for VMI *pitch score* was significant. The effect size of this model was reasonable, as indicated by the amount of the variance explained (adjusted R^2 = .158). Table 2 shows self-rated practice time was the only significant predictor for VMI *pitch score*. The models for VMI *timing score* and the VMI overall score were not significant and explained little variance in the dependent variables (adjusted R^2 < .08). Neither the predictor variables in the timing model nor the overall model reached significance.

Table 2. *Parameter estimates and p-values for the six predictor variables entered into the three multiple regression models with VMI pitch score, VMI timing score and VMI overall score as dependent variables.*

Model					
Predictor	Parameter Estimate	95 % Confidence Intervals		t-value	p-value
		Lower	Upper		
Pitch Score					
Intercept	0.201	-0.306	0.708	0.797	.429
Singing Ability	0.067	-0.014	0.148	1.671	.101
Musicianship	-0.0002	-0.064	0.064	-0.008	.994
Attentive Listening	0.019	-0.027	0.066	0.863	.393
Time Practised Past	0.066	0.005	0.128	2.167	.035
INMI Frequency	0.006	-0.065	0.078	0.190	.850
INMI Length	-0.014	-0.084	0.055	-0.422	.675
Model fit:	Adj. R^2 = .158, $F_{(6, 47)}$ = 2.653, p = .027				
Timing Score					
Intercept	0.397	0.005	0.788	2.038	.047
Singing Ability	0.012	-0.051	0.074	0.375	.709
Musicianship	0.042	-0.007	0.092	1.719	.092
Attentive Listening	-0.017	-0.053	0.019	-0.946	.349
Time Practised Past	0.004	-0.044	0.051	0.160	.874
INMI Frequency	-0.014	-0.069	0.041	-0.507	.615
INMI Length	-0.020	-0.074	0.033	-0.765	.448

Model fit:	Adj. R^2 = -.007, $F_{(6, 47)}$ = 1.065, p = .397				
Overall Score					
Intercept	0.067	-0.375	0.509	0.305	.761
Singing Ability	0.046	-0.024	0.117	1.328	.190
Musicianship	0.038	-0.018	0.094	1.377	.175
Attentive Listening	-0.001	-0.042	0.039	-0.074	.941
Time Practised Past	0.017	-0.036	0.071	0.655	.515
INMI Frequency	-0.019	-0.081	0.043	-0.611	.544
INMI Length	-0.018	-0.078	0.042	-0.594	.555
Model fit:	Adj. R^2 = .074, $F_{(6, 47)}$ = 1.705, p = .141				

In summary, the results indicate that being a practicing musician was the only significant predictor for accuracy in VMI ability, and this association only applied to the pitch task. Musicianship had no association with performance on the VMI timing task or the overall score. Finally, the extent of INMI experienced in daily life had no influence on any of the VMI ability scores, neither in terms of INMI frequency nor length.

In a second analysis stage we assessed the effects of the experimental conditions (i.e. whether the music came back in early/on time/late and flat/same pitch/sharp) and possible interactions with levels of musical practice and INMI activity levels. This analysis, conducted at the level of individual trials, allowed us to take into account the familiarity ratings for each song.

In total there were 587 responses from the 67 participants (16 responses were removed due to individual cases of low song familiarity). Table 3 gives the distribution of responses given by all participants and across all conditions. All experimental conditions were presented equally often across participants, therefore the distribution of responses indicates a clear bias to respond "same pitch" (unaltered) combined with a lesser bias to respond "late timing".

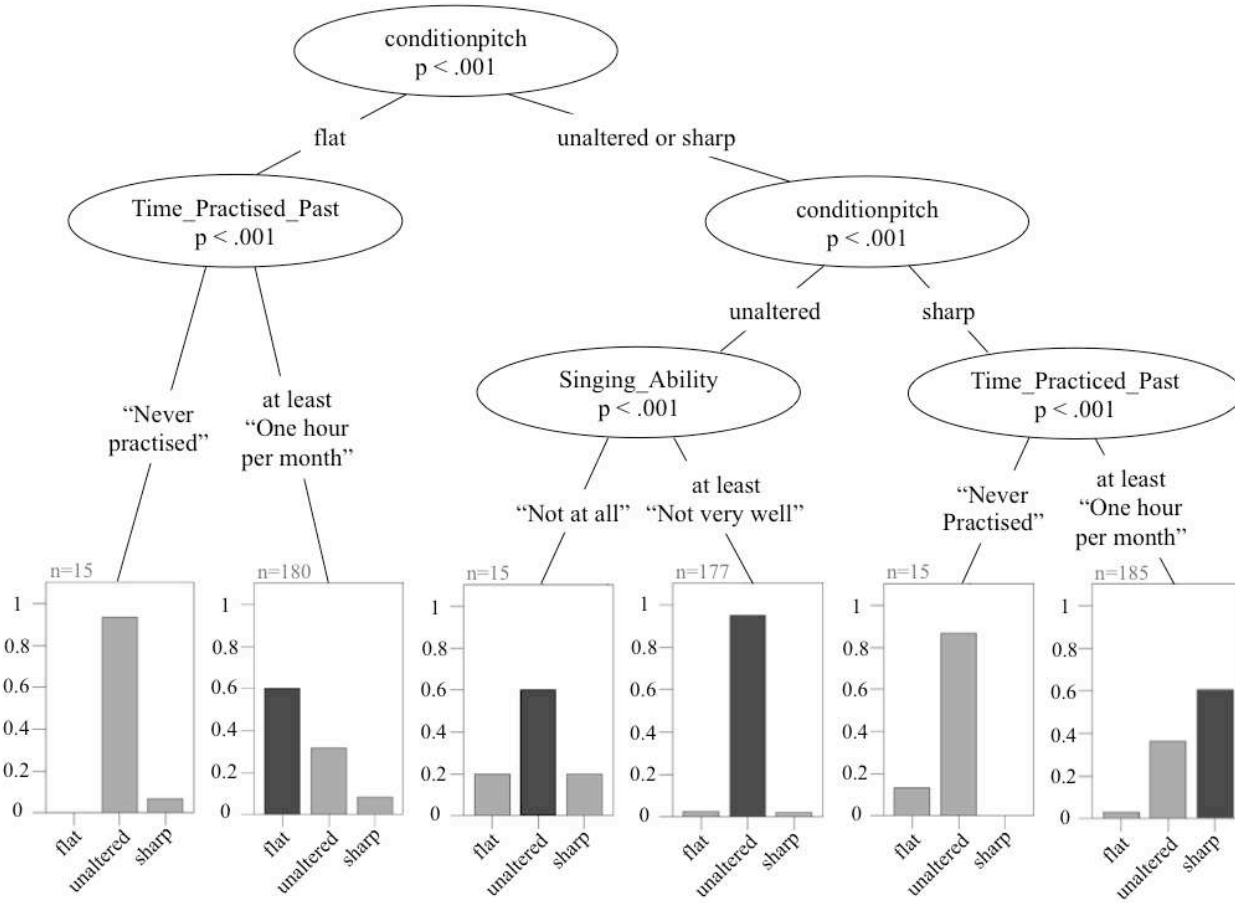
Table 3. Percentages of VMI pitch and timing responses by response category (N=587 observations).

Pitch Responses		
-1 (flat pitch)	0 (unaltered pitch)	1 (sharp pitch)
21.1	55.9	23.0
Timing responses		
-1 (early timing)	0 (unaltered timing)	1 (late timing)
29.0	31.3	39.7

We ran two separate analysis models on the participants' responses, one with pitch and one with timing judgments as the dependent variable. The six predictor variables were used to assess the impact of musical practice and INMI activity levels on both outcomes. In addition, participants' subjective song familiarity ratings and the experimental conditions (pitch and timing) for each trial were added as predictors into the model.

We expected to find high order interactions between variables; therefore we did not use linear regression models. Such models are not well-suited to the task of detecting higher order interactions between predictor variables. Instead we opted for classification tree models, which are an established tool for the identification of significant interactions (Breiman, Friedman, Olshen & Stone, 1984; Müllensiefen, 2009; Strobl, Malley, & Tutz, 2009). Tree models also lend themselves to a graphical interpretation and understanding of the data. We used a family of tree models called conditional inference trees, which combine the rigorous theory of permutation statistics (Hothorn, Hornik, & Zeileis, 2006) with the principle of recursive partitioning (Hothorn, Hornik, & Zeileis, 2008). We used the software package 'party', implemented in the free software environment R. The tree model for the pitch data is given in Fig. 1.

Figure 1. Classification tree model of the pitch responses (N=587). The bar plots in the terminal nodes of the tree represent the distribution of responses for the three different response categories "flat", "unaltered pitch", and "sharp".The correct response is indicated by the dark grey bar in each of the barplots, i.e. the light grey bars indicate the relative amount of incorrect responses. See in-text description for an interpretation of the model.



The model in Figure 1 has a classification accuracy of 72.2%, which is significantly higher ($p < .001$ on a binomial test) than the accuracy expected from a random model (33.3%) and also higher than the accuracy of a model that classifies all observations to the largest response class ('unaltered pitch' responses; 55.9%; $p<.001$ on a binomial test).

The pitch tree model only made use of the experimental conditions, self-rated practice time and self-rated singing ability as predictor variables. For each node of the tree, the p-values indicating the significance of the split are given, as well as a description of the two 'children' (i.e. subgroups) of the split on the independent variable. For the terminal nodes

(i.e. nodes that are not split any further) the distribution of responses is shown in a bar graph with the categories “flat”, “unaltered pitch” and “sharp” responses on the x-axis indicating.

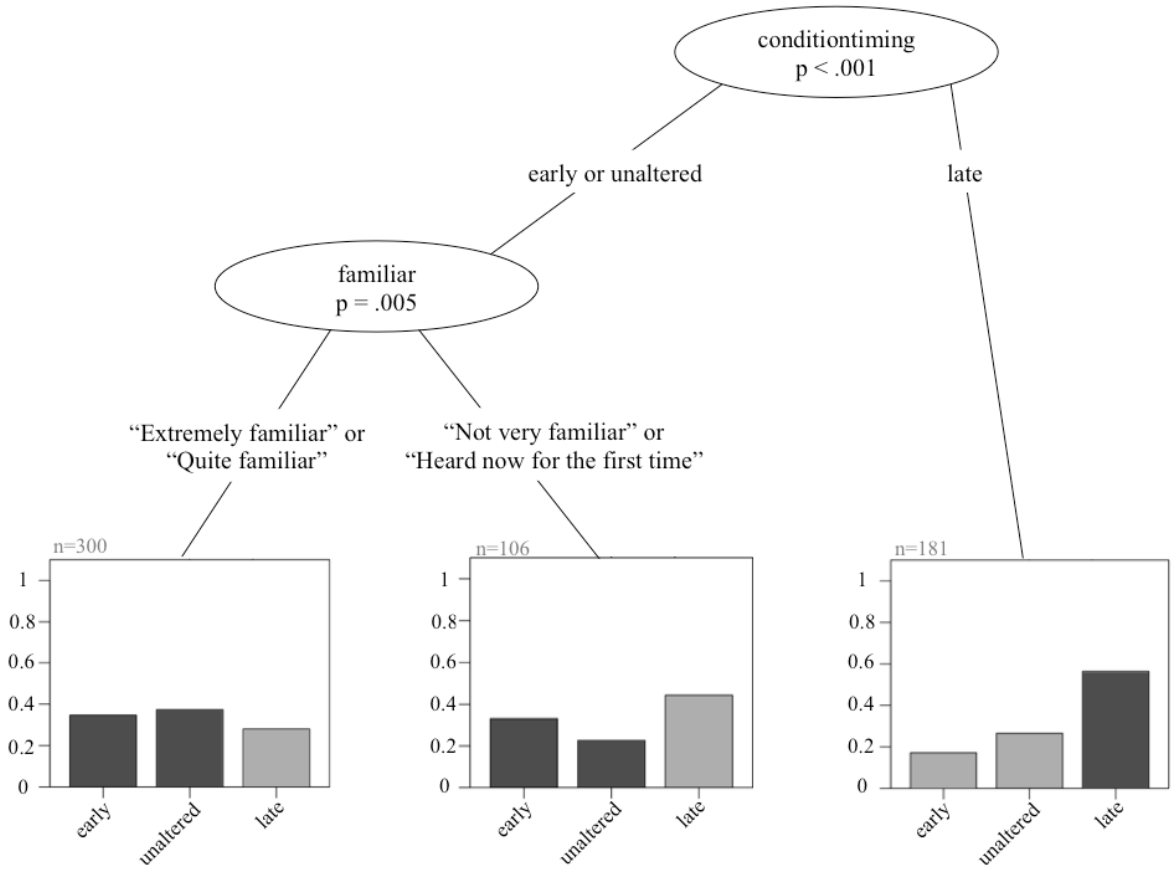
The tree model can be interpreted by starting at the top and following each branch down, to arrive at a final node containing the distribution of responses. For example, follow the first ‘conditionpitch’ node down the ‘flat’ branch (left side of the model), which descends to the left at the ‘time practised past’ node down the ‘Never practised’ branch. This branch can be interpreted as follows: when the music is flat and the participant has never practised an instrument then the response is very likely to be ‘unaltered pitch’. However, when presented with a flat trial a participant with musical expertise (the right branch labelled ‘at least One hour per month’) will have a much higher chance of correctly responding ‘flat’.

Going down the right branch of the top node, the experimental conditions “unaltered pitch” and “sharp pitch” are distinguished. In “unaltered pitch” trials participants with a higher self-rated singing ability identify the unaltered pitch correctly in almost all instances whereas participants with lower self-rated singing abilities are more likely to incorrectly respond “flat” or “sharp”. For experimental trials where the music returns one semitone sharp (right-most branch), the participants with musical experience have a much higher chance of responding “sharp” correctly.

Taken together, the model shows the strong bias for musically untrained participants to respond ‘unaltered pitch’ regardless of the experimental trial, while participants with a minimum of musical training and practice are more able to identify sharp and flat trials correctly on the basis of their VMI. Note that the tree model for pitch responses only reflects interactions between experimental conditions, self-rated singing ability and the amount of time practised in the past: no interaction was found with INMI experience, amount of active musical listening, song familiarity or the experimental timing condition.

A second tree model was created for the VMI timing responses and can be seen in Figure 2. The model has a classification accuracy of 44.4%, which is lower than the pitch tree model but again is significantly higher than the accuracy expected from a random model (33.3%: $p < .001$, binomial test) and the accuracy of a model classifying all observations to the largest response class ('unaltered timing' responses; 39.7%; $p<.011$ on a binomial test). The timing tree model only made use of the experimental timing condition and the participant's subjective familiarity with the song.

Figure 2. Classification tree model of the timing responses (N=587). The bar plots in the terminal nodes of the tree represent the distribution of responses for the three different response categories "early", "unalteredt timing", and "late". See in-text description for an interpretation of the model.



The model in Figure 2 shows that the majority of participants tend to identify late trials correctly (the right-most branch). However, for trials where the music comes in early or at the unaltered timing (the left branch), the response is modulated by the degree of song familiarity. The timing of songs that were not very familiar or heard for the first time tended to be classified as late, while for songs that were extremely or at least quite familiar the response categories were more evenly distributed with 'late' responses being less common. Note that this finding applies regardless of whether the experimental condition was 'early' or 'unaltered timing'. The tendency to incorrectly respond 'late' can therefore be regarded as a bias that is introduced by unfamiliarity with the stimulus.

Discussion

This study explored voluntary musical imagery (VMI) ability and its potential relations to an individual's level of musical training and practice experience or their regular experience of INMI. We predicted an association between musical expertise and VMI accuracy (Aleman et al., 2000; Bishop, Bailes & Dean, 2013), and further hypothesized that a similar improvement in VMI ability may be found in people who experience extreme INMI, due to the common underlying factor of musical imagery repetition in memory.

Accuracy of pitch and timing VMI judgments were tested in a realistic, online pop music listening paradigm. The paradigm assumes that participants use their memory for the song tested to perform the mental imagery task. We found evidence for this assumption from a significant association between song familiarity and accuracy, indicating that participants made use of musical memory when performing the experimental task.

In terms of the main hypotheses we found that, musical practice was positively related to VMI pitch ability, as was expected. This finding is consistent with the VMI pitch results found by Aleman et al. (2000) and is aligned with the findings of increased VMI loudness accuracy in experts reported by Bishop et al., (2013). The present study also extends the work by Aleman et al. (2000) to include a real time VMI pitch listening task, inspired by the studies of Bailes and Bigand (2004) and Kraemer et al. (2005).

This main result implies that the new online VMI paradigm was sensitive enough to detect the expected group expertise effect in VMI ability. Using this same paradigm we found no evidence that people with extreme INMI possessed more accurate VMI pitch ability compared to those with much less extreme INMI experiences. Therefore, we find no evidence

to support the theory that increased involuntary mental repetition of musical imagery in general drives an improvement in VMI ability.

The follow-up analysis of participants' responses at the trial level (tree models) revealed information about the ways musicians' outperformed non-musicians on the VMI pitch task. Participants with no musical training showed a strong bias to respond 'unaltered pitch' in all conditions. By comparison, participants with musical practice experience made more appropriate distinctions between flat, sharp and unaltered pitch levels. Again, this result confirms the findings reported by Aleman et al. (2000) and Bishop et al., (2013) regarding an association between VMI accuracy and musical expertise, but also provides the insight that non-musicians make more mistakes in a VMI task as they are unable to confirm when a pitch shift has occurred.

When interpreting the failure of non-musicians to detect pitch modifications in the VMI task, it is worth noting that the VMI pitch task comprises at least two separate components; the ability to imagine the continuation of the song, drawing on memory resources, as well as a mental comparison between the imagined and heard pitch. From the present data it is not possible to infer whether it is the imagination, memory or the comparison component or another factor that is associated with poorer performance in non-musicians compared to musicians.

In addition to musical practice, self-rated singing ability was associated with improved performance on the VMI pitch task, specifically to an increased ability to identify "correct" pitch trials. Singing is a voluntary musical production activity that is closely related to the corresponding VMI and which need not relate to musical practice in a formal sense (Halpern, 1989; Zatorre & Halpern, 2005; Zatorre, Halpern & Bouffard, 2010). The positive

association between singing and performance on the present VMI pitch task corroborates this relationship.

The present study did not identify any variables associated with overall VMI timing accuracy of the participants. VMI timing performance was worse than that on the VMI pitch task. These findings are consistent with Janata and Paroo (2006), who suggested that temporal images are more susceptible to distortion than pitch images, and with the finding that the neural substrates differ for temporal and pitch judgments (Hyde & Peretz, 2004). However, it is worth noting that we are not able to claim that the timing manipulation in the present paradigm (time shift by 2 beats) is of equivalent difficulty level to the pitch manipulation (pitch shift by 1 semitone). For a follow-up experiment it would be useful to equate difficulty levels for both response tasks in a prior calibration experiment, using an item response approach (de Ayala, 2009).

The tree model on the VMI timing data at trial level confirmed that neither musical practice nor singing ability, nor any other indicator of musical activity, had any impact on performance on the timing task. This result is somewhat contrary to findings in the beat perception literature, where a positive relationship has been noted between musical expertise and beat perception abilities (Repp, 2010). However, the present VMI timing task is of a different nature to the beat perception tasks used in previous studies. Hence, studies that compare individual differences on comparable timing and beat perception tasks, with and without imagery components, would be necessary to identify where the timing advantages of musical expertise are to be located.

Overall, the present study offers no evidence for the hypothesis that musical practice activity and INMI have a similar 'training' effect on VMI ability by virtue of increasing the repetition of musical imagery in memory. This conclusion does not rule out the possibility

that individual differences in INMI experiences may relate to differences in musical memory. Research has established that musicians have a superior memory for many aspects of musical sounds compared to non-musicians (for reviews see Schulze & Koelsch, 2012; Williamson, Baddeley & Hitch, 2010). Therefore, improved musical memory could be one factor that drives the positive association between musical practice and VMI pitch ability in musicians. Future research investigating the relationship between INMI and memory could test whether people who experience extreme INMI (regardless of musical background) have a better short-term or long-term memory for music, either familiar or novel, compared to people who rarely experience INMI.

One limitation of the present study is the difficulty of assessing the extent of musical activity and INMI that participants experience. The present study utilized questionnaires that were available at the time of testing. An improved approach for the future would be to use new, valid and reliable self-report instruments (e.g. the Gold-MSI, Müllensiefen et al., 2014b and the IMIS, Floridou et al., this issue) to obtain more precise measures of musical engagement and quantifiable aspects of INMI experiences. Another concern with INMI measurement in the present study is the reliance on recent reports of activity compared to lifetime assessment of experiences. Future studies could aim for increased control across the conditions by using INMI report methods akin to those employed in music expertise research, such as asking participants to report on long-term patterns of INMI activity, or testing VMI ability at different life stages (for a review see Strait & Kraus, 2014).

In summary, the present study comprised a new experimental paradigm that examined VMI pitch and timing abilities using highly familiar pop music. We found no differences in VMI ability in pitch or timing between people that experience extreme INMI

and people that experience infrequent, short INMI. Our conclusion is that while active musical practice is related to improved VMI pitch ability, extreme INMI is not.

References

Aleman, A., Nieuwenstein, M. R., Bocker, K. B. E., & de Haan, E. H. F. (2000). Music training and mental imagery ability. *Neuropsychologia, 38(12),* 1664–68.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In Spence, K. W., & Spence, J. T. (Eds.), *The psychology of learning and motivation (Vol. 2, pp. 89–195).* New York: Academic Press.

De Ayala, R. J. (2009). *Theory and practice of item response theory*. Guilford Publications.

Baddeley, A .D., & Hitch, G. (1974). Working memory. In G.H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory (Vol. 8, pp. 47–89).* New York: Academic Press.

Baddeley, A. D., & Logie, R. H. (1992). Auditory imagery and working memory. In D. Reisberg (Ed.)*, Auditory Imagery (pp. 179 –197)*. Hillsdale, NJ: Erlbaum.

Baddeley, A. D., & Andrade, J. (2000). Working memory and the vividness of imagery. *Journal of Experimental Psychology: General, 129,* 126-145*.*

Bailes, F. & Bigand, E. (2004). A tracking study of mental imagery for popular classical music. *Paper presented at 8th International Conference on Music Perception and Cognition.*

Bailes, F. (2007). The prevalence and nature of imagined music in the everyday lives of musical students. *Psychology of Music, 35,* 1-16*.*

Bailes, F., Bishop, L., Stevens, C. J., & Dean, R. T. (2012). Mental imagery for musical changes in loudness. *Frontiers in Perception Science, 3,* 525*.*

Beaman, C. P., & Williams, T. I. (2010). Earworms ('stuck song syndrome'): Towards a natural history of intrusive thoughts. *British Journal of Psychology, 101(4),* 637-653*.*

Bishop, L., Bailes, F., & Dean, R. T. (2013). Musical expertise and the ability to imagine loudness. *PloS one*, *8*(2), e56052.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Brodsky, W., Kessler, Y., Rubinstein, B. S., Ginsborg, J., & Henik, A. (2008). The mental representation of music notation: Notational audiation. *Journal of Experimental Psychology: Human Perception and Performance, 34,* 427-445.

Christoff, K., Ream, J. M., & Gabrieli, J. D. E. (2004). Neural basis of spontaneous thought processes. *Cortex, 40(4-5*), 623–630.

Darwin, C. (1972). An auditory analogue of the Sperling partial report procedure: Evidence for a brief auditory storage. *Cognitive Psychology, 3(2),* 255–67.

Floridou, G., Williamson, V., Stewart, L., & Müllensiefen, D. (this issue). The Involuntary Musical Imagery Scale (IMIS). *Psychomusicology: Music, Mind and Brain.*

Funt, B. V. (1980). Problem-Solving with Diagrammatic Representations. *Artificial Intelligence 13(3)*, 201-230.

Halpern, A. R. (1988a). Mental scanning in auditory imagery for songs. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14(3)*, 434.

Halpern, A. R. (1988b). Perceived And Imagined Tempos Of Familiar Songs. *Music Perception, 6(2)*, 193-202.

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

Halpern, J. (1992). Effects of historical and analytical teaching approaches on music

appreciation. *Journal of Research in Music Education*, *40(1)*, 39-46.

Halpern, A. R., & Bartlett, J. C. (2011). The Persistence Of Musical Memories: A Descriptive Study Of Earworms. *Music Perception, 28(4),* 425-432.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional framework. *Journal of Computational and Graphical Statistics, 15(3),* 651-674.

Hothorn, T., Hornik, K., & Zeileis, A. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*, 492–514.

Hyde, K. & Peretz, I. (2004). Brains that are out of tune but in time. *Psychological Science, 15(5),* 356-360.

Hyman, I. E. Jr, Burland, N. K., Duskin, H. M., Cook, M. C., Roy, C. M., et al. (2013). Going Gaga: Investigating, creating, and manipulating the song stuck in my head. *Applied Cognitive Psychology, 27,* 204–215.

Janata, P., & Paroo, K. (2006). Acuity of auditory images in pitch and time. *Perception & Psychophysics, 68(5),* 829-844.

Kosslyn, S. M. (1994). *Image and brain: The resolution of the imagery debate*. Cambridge, MA: MIT Press.

Kraemer, D. J., Macrae, C. N., Green, A. E., & Kelley, W. M. (2005). Musical imagery: sound of silence activates auditory cortex. *Nature*, *434(7030)*, 158-158.

Liikkanen, L. A. (2008). Music in everymind: Commonality of involuntary musical imagery. In K. Miyazaki, Y. Hiragi, M. Adachi, Y. Nakajima, & M. Tsuzaki (Eds.), *Proceedings of the 10th International Conference on Music Perception and Cognition (ICMPC10)* (pp. 408–412).

Liikkanen, L. A. (2012). Musical activities predispose to involuntary musical imagery. *Psychology of Music, 40(2),* 236-256.

Lucas, B. J., Schubert, E., & Halpern, A. R. (2010). Perception of emotion in sounded and imagined music. *Music Perception*, *27*, 399–412.

Mason, M. F., Norton, M. I., Van Horn, J. D., Wegner, D. M., Grafton, S. T., & Macrae, C. N. (2007). Response to comment on "Wandering mind's: The default network and stimulus-independent thought". *Science, 317,* 5834.

Müllensiefen, D. (2009). Statistical techniques in music psychology: An update. In: R. Bader, C. Neuhaus, C. Morgenstern (eds.), *Concepts, Experiments, and Fieldwork: Studies in Systematic Musicology* (pp. 193-215). Frankfurt a.M.: Peter Lang.

Müllensiefen, D., Fry, J., Jones, R., Jilka, S. R., Stewart, L. & Williamson, V. J. (2014a). Individual differences in spontaneous involuntary musical imagery. *Music Perception, 31(4),* 323-335.

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014b). The Musicality of Non-Musicians: An Index for Assessing Musical Sophistication in the General Population. *PloS One , 9(2),* e89642.

Pylyshyn, Z. W. (2002). Mental Imagery: In search of a theory. *Behavioral and Brain Sciences, 25(2),* 157-237.

Raichle, M. E., & Snyder, A. Z. (2007). A default mode of brain function: a brief history of an evolving idea. *NeuroImage, 37(4),* 1083-1090.

Reisberg, D. (1992). *Auditory Imagery*. Hillsdall, N. J.: Erlbaum.

Repp, B. H. (2010). Sensorimotor synchronization and perception of timing: Effects of music training and task experience. *Human Movement Science, 29*, *200-213.*

Schulze, K., & Koelsch, S. (2012). Working memory for speech and music. *Annals of the New York Academy of Sciences*, *1252*, 229-236.

Shepard, R., & Metzler. J. (1971). Mental rotation of three dimensional objects. *Science*, *171(972)*, 701-3.

Smith, J. D., Reisberg, D., & Wilson, M. (1992). Subvocalization and auditory imagery: Interactions between the inner ear and inner voice. In D. Reisberg (Ed.), *Auditory imagery* (pp. 95-119). Hillsdale, NJ: Erlbaum.

Smith, J. D., Wilson, M., & Reisberg, D. (1995). The Role Of Subvocalization In Auditory Imagery. *Neuropsychologia*, *33(11)*, 1433-54.

Strait, D.L., & Kraus, N. (2014) Biological impact of auditory expertise across the life span: musicians as a model of auditory learning. *Hearing Research, 308*, 109-121.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods, 14(4)*, 323.

The Official Charts Company | The UK Charts | Top 40. (n.d.). Retrieved 2014, from http://www.officialcharts.com/

Webber, R. J., & Brown, S. (1986). Musical imagery. *Music Perception*, *3*, 411– 426.

Williamson, V.J & Jilka, S.R (2014) Experiencing earworms: An interview study of Involuntary Musical Imagery. *Psychology of Music*, *42(5)*, 653-670.

Williamson, V. J., Liikkanen, L.A, Jakubowski, K., & Stewart, L. (2014) Sticky Tunes: How do people react to involuntary musical imagery? *PLOS ONE*, *9(1)*, e86170

Williamson, V. J., Jilka, S. R., Fry, J., Finkel, S., Müllensiefen, D., & Stewart, L. (2012). How do "earworms" start? Classifying the everyday circumstances of Involuntary Musical Imagery. *Psychology of Music*, *40(3)*, 259-284.

Williamson, V. J., Baddeley, A. D., & Hitch, G. J. (2010). Musicians' and nonmusicians' short-term memory for verbal and musical sequences: Comparing phonological similarity and pitch proximity. *Memory and Cognition*, *38(2)*, 163-175.

Zatorre, R. J. & Halpern, A. R. (2005). Mental concerts: Musical imagery and auditory cortex. *Neuron*, *47*, 9-12.

Zatorre, R. J., Halpern, A. R., & Bouffard, M. (2010). Mental reversal of imagined melodies: A role for the posterior parietal cortex. *Journal of Cognitive Neuroscience*, *22*, 775-789.

Zatorre, R. J., Halpern, A. R., Perry, D. W., Meyer, E., & Evans, A. C. (1996). Hearing in the mind's ear: A PET investigation of musical imagery and perception. *Journal of Cognitive Neuroscience*, *8(1)*, 29-46.

Appendix 1. Songs used in imagery experiment, with edit times.

Song title	Artist	Start (Seconds)	Dropout (Seconds)	Return (Seconds)	End (Seconds)
Hound Dog	Elvis Presley	0	19.5	28.4	37.8
I wanna hold your hand	The Beatles	0	11.4	19.0	30.8
I'm a believer	The Monkees	4.8	17.1	27.3	35.8
Kung fu fighting	Carl Douglas	49.9	70.7	81.5	89.6
What a wonderful world	Louis Armstrong	22.7	57.4	68.6	76.8
Beat it	Michael Jackson	14.4	31.7	41.2	48.4
Baby one more time	Britney Spears	20.5	58.0	70.5	84.1
Hey ya	Outkast	0	26.2	34.4	46.8
I gotta feeling	Black Eyed Peas	15.2	33.4	44.8	56.4