ASSESSING THE EFFECTIVENESS OF DIFFERENT TRAINING PARADIGMS IN PITCH CATEGORISATION

by

André Grenier

Department of Psychology

Submitted in Partial Fulfilment

of the requirements for the degree of

Bachelor of Arts

in

Honours Psychology

Faculty of Arts and Social Science

Huron University College

London, Canada

April 30, 2023

HURON UNIVERSITY COLLEGE

CERTIFICATE OF EXAMINATION

Advisor:	Dr. Stephen Van Hedger
Reader:	Dr. Christine Tsang

The thesis by:

André Grenier

entitled:

Assessing the Effectiveness of Different Training Paradigms in Pitch Categorisation Training

is accepted in partial fulfilment of the requirements for the degree of

Bachelor of Arts

in

Honours Psychology

April 30, 2023	Dr. Christine Tsang
Date	Chair of Department

Abstract

Gamification - the process of incorporating game-like elements (e.g., points, levels) into traditionally non-game domains - is a concept that has been increasingly commonplace over the past decade. In cognitive psychology, gamification has begun to be incorporated into research paradigms, particularly in the context of learning (e.g., language or music learning). Although research in this area is still nascent, initial results suggest that gamification may lead to significantly accelerated learning trajectories and improved generalisation of learning compared to more traditional paradigms, particularly in challenging domains (such as learning non-native speech sounds). The present experiment extends this gamification research to assess learning efficacy of absolute pitch (AP) – a notoriously difficult learning context in which listeners must categorise pitched sounds based on musical pitch class. We compared a gamified paradigm, in which listeners moved a spaceship on the screen to navigate through different "gates" associated with musical notes, to a more traditional paradigm, used in prior AP research, to assess performance for both the trained sounds (assessed via a "rote test") and novel sounds that belonged to the same trained categories (assessed via a "generalisation test"). While the gamified paradigm led to significant learning in the rote test, performance in the generalisation test was on marginally above chance. In contrast, and contrary to the study's predictions, the standard paradigm led to robust learning in both the rote and generalisation tests, with learning rates exceeding what has been reported in past studies. We offer that gamification allows us the vocabulary and concepts necessary to explain this discrepancy.

Keywords: *absolute pitch, videogames, perceptual learning, music cognition, gamification, pitch categorisation*

Acknowledgements

Words cannot express my immense gratitude to my thesis advisor Dr. Stephen Van Hedger. This thesis would not have been possible without his incredible care, dedication, and passion. The many conversations and hours spent working together helped give me confidence in my academic abilities. The effects of this bled into each of my endeavours. I am truly honoured and privileged to have the opportunity to work with you.

I would like to extend my appreciation to the chair of the psychology department at Huron, Dr. Christine Tsang (who was also my second reader), for giving me the opportunity to enroll in the thesis program a year early. Her interest in my thesis was very affirming, and I would like to thank her for the excellent organisation of the annual CURL conference. Special thanks should also be given to all the professors of psychology that helped and shaped me along my journey. Their collective enthusiasm has been a beacon of inspiration. I would be remiss not to mention the support of my fellow members of the Huron Auditory Perception laboratory. Being able to share our experiences throughout this process was very valuable.

My success must also be attributed to my wonderful family: my parents, brothers and grandparents. The support and stability they have offered throughout my entire life has been invaluable, especially these past few years. I would also like to thank my dog, Bisou, for being a great source of comfort and a wonderful listener.

Table of Contents

CERTIFICATE OF EXAMINATION ii

Abstract iii

Acknowledgements iv

Table of Contents v

Introduction 1

Method 6

Participants 6

Materials 7

Procedure 8

Calibration 8

Videogame 9

Control 10

Questionnaire and Debriefing 12

Results 13

Assessing Accuracy against Chance Performance 13

Assessing Accuracy across Condition and Test 13

Self-Reported Engagement 15

Correlating AP Learning with Additional Variables 16

Discussion 16

References.. 25

Curriculum Vitae 30

Introduction

Auditory pitch is a critical dimension for the perception and understanding of Western music (e.g., Kuusi, 2009). However, there is a large amount of individual variability in how pitch is perceived and remembered in the population (Peretz et al., 2003). As such, it is perhaps not surprising that research examining how the perception and memory of musical pitch can be improved via training has been an integral part of the field of music perception and cognition for several decades (Hartman, 1954; Cuddy, 1968; Mull, 1925). However, these training paradigms have been inconsistently applied across studies and are often repetitive and tedious. Thus, the present research study examines how a game-based training paradigm, relative to a standard psychological training paradigm, influences the learning of absolute pitch categories in music.

Absolute pitch (AP), also known as perfect pitch, is the ability to name or produce a note in the absence of a reference note (e.g., see Takeuchi & Hulse, 1993 for a review). Although AP may hinder some aspects of musical processing, such as relative pitch processing (e.g., see Miyazaki, 1993), AP is generally viewed as a desirable musical trait (Deutsch, 2002) and affords an individual significant advantages in musical tasks such as aural dictation (Dooley & Deutsch, 2010). Consequently, understanding how AP is acquired is both of general and scientific interest. AP has been described as a model system for understanding gene-by-environment interactions (e.g., Zatorre, 2003) and is thus an important phenomenon to study in understanding the development and maintenance of complex behaviours.

Some of the greatest puzzles surrounding AP have to do with its presumed rarity and how it is acquired. Absolute pitch is only seen in around 0.01% of the population (Bachem, 1955). This low base rate in the general population is surprising when one considers that the process of assigning labels to pitched sounds is akin to assigning absolute labels to other sounds (e.g., vowels) and to other perceptual experiences (e.g., colour), which most individuals do with ease. In fact, individuals without AP have been likened to individuals with *colour anomia* in the visual realm – a phenomenon in which individuals can perceive differences in colour (e.g., discriminating wavelengths of light corresponding to red versus green) but cannot ascribe labels to these perceptual experiences (Levitin & Rogers, 2005). Given the seeming simplicity of AP (ascribing a category label to a pitched sound), combined with its desirability in musical contexts, it is perhaps not surprising that there has been substantial interest in the trainability of AP across the lifespan.

Although the mechanisms underlying the acquisition of AP are currently debated, there is a great deal of skepticism around the possibility of teaching absolute pitch to adults. The dominant views in the literature are that absolute pitch is either an ability present since birth (Baharloo, 2001; Theusch, 2010) or developed within a specific critical period in childhood (Russo et al., 2003). These views were supported in principle by the observation that no adult AP training study yielded performance comparable to a "genuine" AP possessor. However, many of these previous studies reporting null results were heterogeneous in training approach and training duration. As such, it is difficult to strongly interpret these null results as "evidence of absence" – i.e., that AP cannot be trained in adulthood. These null results have also recently been challenged by studies demonstrating successful AP training in a subset of adults (Van Hedger et al., 2019; Wong, Lui et al., 2020; Wong, Ngan et al., 2020). However, these more recent findings are based on training paradigms that were lengthy (consisting of hours of training, spaced out over dozens of training sessions) and tedious (consisting of sometimes thousands of repetitive perceptual decisions). Nevertheless, they provide important proof-of-concept findings that AP is trainable into adulthood, with some participants demonstrating sufficient speed and accuracy of classifying notes that is behaviourally indistinguishable from “genuine” AP possessors.

The current state of AP training is thus conflicted. The variability in training approaches, length of training, and number of participants reported in each study makes it difficult to draw generalised conclusions about the relative extent to which AP is trainable in adulthood. However, given the debate about whether AP is even possible to acquire in adulthood, it is important to consider the factors that would optimise performance in an AP training paradigm. Many existing approaches have used training paradigms that are well controlled but tedious for participants, which might lower an individual’s motivation to perform well and thus might provide an inaccurate assessment of the trainability of AP. This problem is not unique to absolute pitch training. In fact, most training paradigms assessing perceptual learning are lengthy and repetitive, often consisting of thousands of judgments of highly similar stimuli (Dosher & Lu, 1998). This is reflective of a larger goal of experimental research to create an environment in which the effects of independent variables are as clearly differentiated as possible. Although the addition of engaging elements to a study might improve participant enjoyment and might increase ecological validity, it also introduces potential confounding variables that make it difficult to isolate individual elements of what makes a perceptual training program effective.

However, this trade-off between ecological validity and experimental control is not *inherently* a problem, especially if the underlying research question seeks to evaluate the upper limits of what is possible via a training paradigm (rather than deconstructing how specific factors relate to training efficacy). Nacke and Deterding (2017) explain that before theories can be formed, evidence must first establish base validity (e.g., whether adults can learn AP). Given the controversy of whether AP is even trainable in adulthood, it serves as a particularly promising model system for assessing how the development of an interactive and engaging training paradigm, modeled after elements of videogames, relates to learning efficacy.

Researchers have been interested in videogames for a long time. This of course means that videogames are well studied. However, this has given rise to complicated terms and definitions. The two terms that need further clarification for the purposes of the present study are *serious games* and *gamification*. Serious games are normally defined as games created with “an explicit and carefully thought-out educational purpose and are not intended to be played primarily for amusement” (Laamarti et al., 2014, p.1). Examples of serious games can be seen in the recent efforts in creating videogames to introduce children to coding, reading, and mathematics. Gamification, on the other hand, is normally seen as “a process of enhancing services with (motivational) affordances in order to invoke gameful experiences and further behavioral outcomes” (Hamari et al., 2014, p.2). Gamification is commonly seen in consumer contexts, such as the point rewards systems at fast food or coffee places (e.g., earning free items from the menu after many purchases) (Nacke & Deterding, 2017). These two concepts, however, are not mutually exclusive. Indeed, the concept of *game-based learning* represents elements of both serious games and gamification. Although never given a proper definition by the literature from which it is taken, game-based learning is the utilisation of serious games for experimental research or the gamification of traditional training paradigms. Due to the nature of game-based learning including such a large degree of variables, it could very well be used to test whether certain training effects even exist. Future research can then narrow the scope of variables to identify what causes these effects to occur.

Game-based paradigms have been increasingly popular in research and continue to be explored in current research. In the field of music cognition, game-based learning has been employed for improving singing (Paney & Kay, 2015), as well as a means of improving pitch perception conceptually similar to absolute pitch (Yang & Cheng, 2020; Li & Gu, 2021). Importantly, the benefits of these game-based models extend beyond the performance gains observed in the domain of learning (e.g., singing accuracy). The enjoyment that participants report experiencing with these games is much higher than traditional training paradigms (Cheng et al., 2015). However, these recent approaches to game-based learning are not particularly accessible, given their technological requirements. For example, Paney and Kay (2015), as well as Li and Gu (2021), both utilise game-based models that require high quality microphones to be played. This presents an issue, as higher quality microphones do not come built into existing computers and can be very expensive. Other studies, such as Yang and Cheng (2020), employ cutting-edge technology in virtual reality. This further inhibits the accessibility beyond the price, as the installation and implementation require significant space and additional resources (e.g., electrical outlets). Although the pursuit of methodologies utilising more advanced technology can have positive implications, the present study takes a simpler approach which is geared towards maximising accessibility.

The inspiration for this study comes from an influential study on teaching native pronunciation to foreign speakers by Lim and Holt (2011). The game-based paradigm used by Lim and Holt (2011) was adapted from an older study (Wade & Holt, 2005), with the only change being the to-be-learned auditory stimuli. What makes the Lim and Holt (2011) study so notable is the accelerated learning trajectories in its game-based learning environment. The authors found that two and a half hours using their game-based model resulted in learning comparable to 45 hours of training with a traditional paradigm. This provides further validation towards a simpler game design when one considers that for this game-based paradigm, the only requirements were a computer and headphones. This study is the ideal case of having very high results while being very accessible. Learning non-native pronunciations is also very difficult, a factor which is believed to be important for maximal engagement in videogames (Hamari et al., 2016). By extension, AP should also be a well-suited task for game-based learning, given its difficulty of acquisition discussed earlier.

The goal of the present study was to create a simple yet effective game-based learning paradigm to train AP categories in participants, conceptually extending the work of Lim and Holt (2011) to a musical domain. Aesthetic and narrative appeal are important in developing a game-based learning paradigm, and thus the present study uses a space theme (borrowing heavily from retro-videogame aesthetics, specifically “Galaga”). To directly assess the efficacy of game-based learning in the present context, we include a traditional training paradigm (in which the learning environment is more controlled, yet training is highly repetitive) as a control condition. The control condition will be modeled after an influential AP training paradigm (Gervain et al., 2013). Based on prior research in game-based learning, we hypothesise that (1) participants in the game-based learning condition will show enhanced learning compared to participants in the control condition, and (2) participants in the game-based learning condition will report that the learning experience was more engaging than participants in the control condition.

Method

Participants

100 participants from Amazon Mechanical Turk were recruited (videogame condition: *n* = 50, control condition: *n* = 50), and 92 were included in the main analyses (*M* = 38.57 years, *SD* = 9.28 years, range = 22 to 72 years). Two participants’ data from the videogame condition were not saved to our server, and an additional six participants (videogame condition: n = 2, control

condition: n = 4) self-reported possessing AP, meaning 46 participants were included in the videogame condition and 46 participants were included in the control condition. Recruitment used CloudResearch, which interfaces with Mechanical Turk and allows additional participant recruitment parameters. Participants had at least 90% approval rating from prior assignments, with a minimum of 50 previously completed assignments. Participants were "CloudResearch Approved", meaning all participants had passed internal attention checks administered by CloudResearch. Participants were paid $8 USD for completing the study.

Materials

Images for the videogame condition were taken from unsplash.com and pixilart.com. The remaining images were created using Photoshop. Sounds used in training and rote testing across both conditions were generated using sampled instruments in Reason 4 (Propellerhead: Stockholm, Sweden). There was only one sound for each of the four trained categories (C, D#, F#, A). However, each sound contained multiple instruments playing the to-be-learned note across multiple octaves (e.g., a piano and a synthesizer playing the same note simultaneously). In the videogame condition, the sounds were longer with trials of a fixed duration (as the gates moved at a pre-specified pace). Therefore, each sound file was 8000 milliseconds (ms) in duration and had an isochronous rhythm (the to-be-learned note repeated 32 times in the span of the 8000 ms sound). The control condition was a more standard training paradigm (participants heard a sound and then made a response). For this reason, the control condition sounds were 1000 ms in duration, as their length was not tied to a visual event on the screen. The control condition sounds contained the same multi-instrument, multi-octave instrumentation as the videogame sounds, but did not have a rhythmic element (i.e., the to-be-learned note was played once and sustained over the 1000 ms duration). The sounds used for assessing generalisation of learning were Shepard tones generated in Matlab (MathWorks: Natick, MA). Three octaves of Shepard tones were stacked upon one another (e.g., C3, C4, and C5 for the note C). Each Shepard tone consisted of seven harmonics (three below the fundamental frequency and three above the fundamental frequency). Shepard tones have clear pitch chroma but are ambiguous in pitch height (e.g., see Shepard, 1964), making them ideal for disentangling learning of pitch chroma and pitch height. Additionally, the use of layered Shepard tones has been used in prior work to effectively create an illusion of continually ascending or descending musical scales (Cormier, 2021), suggesting that these tones are indeed ambiguous with respect to pitch height. Like the training sounds, the Shepard tones for the videogame condition were 8000 ms in duration and repeated the note 32 times over the duration of the trial event). In contrast, the Shepard tones in the control condition were 1000 ms in duration and did not repeat the note within a trial. All sounds were digitised at 44.1 kHz with 16-bit depth and root mean square normalised to an average amplitude of -25 dB relative to full scale. The experiment was programmed in jsPsych 7 (de Leeuw, 2015)

Procedure

Participants were recruited to each condition on consecutive days on Mechanical Turk. The videogame condition was run first, followed by the control condition the following day. Participants who completed the videogame condition were ineligible to complete the control condition.

Calibration

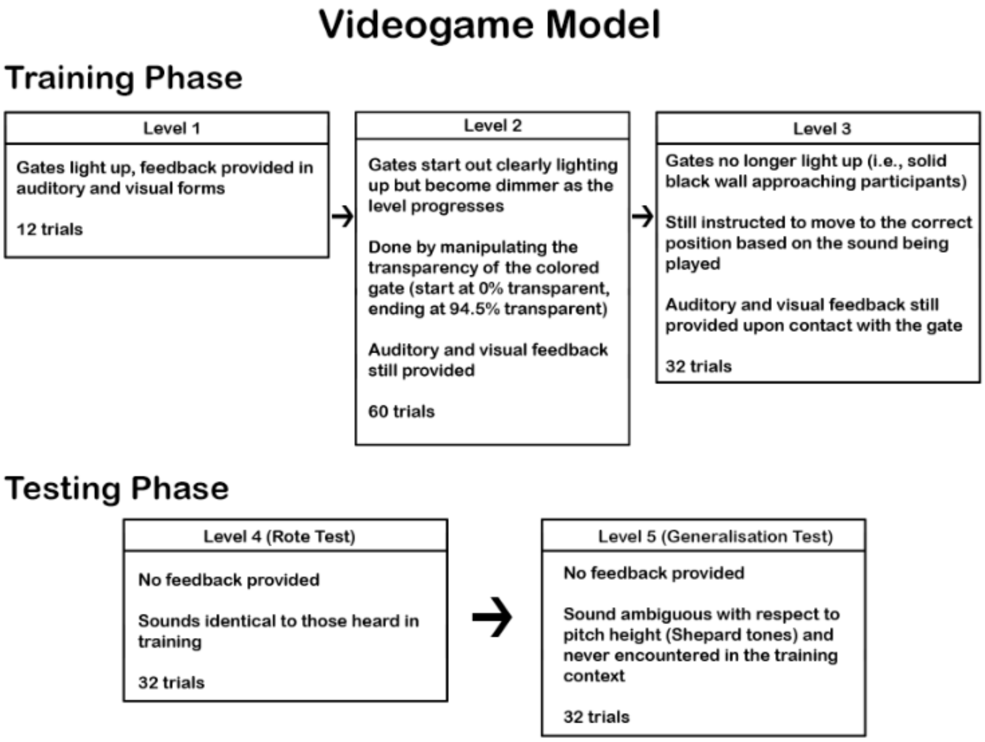
All participants, regardless of condition, completed a short auditory calibration prior to the main AP training and testing task. Participants were presented with a sample of music, normalised to the same level as the sounds in the main AP task, and asked to adjust their volume to a comfortable level. They were then asked whether they were wearing headphones or earbuds. Regardless of whether they answered “yes” or “no”, they then completed a performance-based headphone/earbud assessment. This assessment, based on Milne et al. (2021), consisted of six trials, each trial consisting of three bursts of noise. Of these three bursts of noise, one of them contained a “Huggins Pitch” - a whistling tone that can only be perceived when there is clear audio channel separation, as is the case when wearing headphones. Participants made a forced-choice judgment as to which noise burst contained the tone, and performance of at least five correct responses out of six was used to indicate headphone use. A total of 80.4% of participants reported wearing headphones, and 67.4% passed the performance-based measure of headphone use. Given that headphone use was recommended but not required, neither the self-reported headphone use, nor the performance-based headphone measure were used as exclusion criteria.

Videogame

Participants in the videogame condition played a space-themed game. In this game, participants controlled a spaceship trying to escape an exploding planet. Participants were told the rules and story of the game by a little purple character named Tommy. Throughout the game, Tommy would prepare participants for the increases in difficulty that accompanied “level” changes (i.e., experimental blocks). Sometimes this was through a hint (e.g., “*Pay attention to the sound passing through each gate, it may become important later*” to foreshadow the importance of learning the sound/gate mappings) or by directly telling them what has changed about the game (e.g.: “*Oh no, the lights are out! We’re in total darkness!*” to refer to the fact that the coloured gates would no longer light up and thus participants would have to rely solely on the sounds to navigate). The general mechanics of the game relied on moving a spaceship along four predetermined tracks to pass through gates. Participants controlled the spaceship by pressing either the up and down arrow keys or the “w” and “s” keys to move up or down one track. The gates participants had to move through were each assigned a different note (C, D#, F#, A) and emitted their unique note to indicate which gate participants needed to pass through. Building on these general mechanics, participants played through five levels of increasing difficulty. Figure 1 provides an overview of the progression of levels for the videogame condition.

Figure 1

Overview of the procedure for the videogame condition



Videogame Model

Training Phase

Level 1

Gates light up, feedback provided in auditory and visual forms

12 trials

Level 2

Gates start out clearly lighting up but become dimmer as the level progresses

Done by manipulating the transparency of the colored gate (start at 0% transparent, ending at 94.5% transparent)

Auditory and visual feedback still provided

60 trials

Level 3

Gates no longer light up (i.e., solid black wall approaching participants)

Still instructed to move to the correct position based on the sound being played

Auditory and visual feedback still provided upon contact with the gate

32 trials

Testing Phase

Level 4 (Rote Test)

No feedback provided

Sounds identical to those heard in training

32 trials

Level 5 (Generalisation Test)

No feedback provided

Sound ambiguous with respect to pitch height (Shepard tones) and never encountered in the training context

32 trials

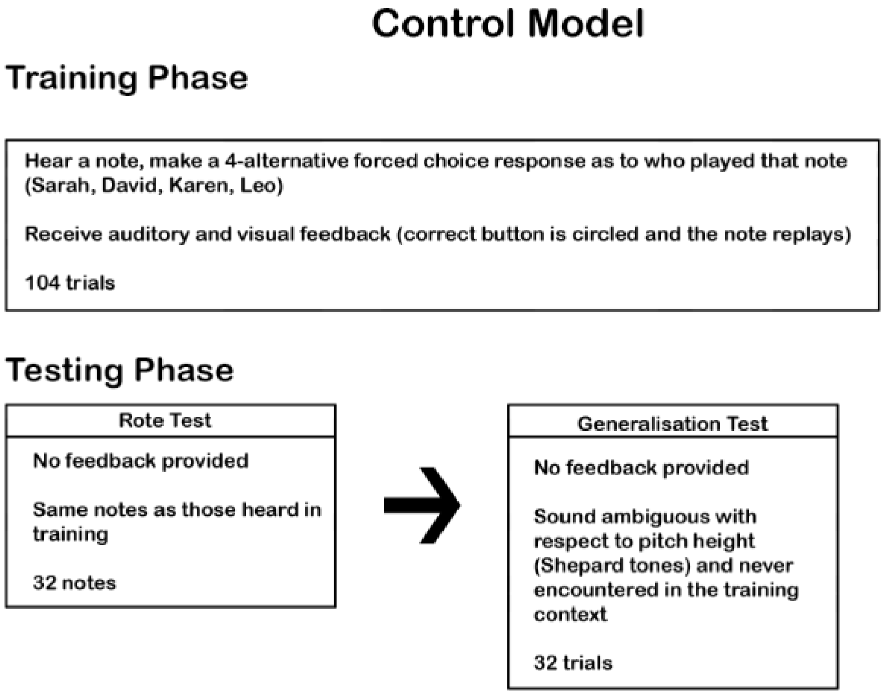
Control

The control condition was based on standard training and testing approaches used in AP research; with a particular alignment to the paradigm used in Gervain et al., (2013). The control paradigm assigned arbitrary names to the four trained notes (the note “C” was given the name “Sarah,” the note “D#” was given the name “David,” the note “F#” was given the name “Karen,” and the note “A” was given the name “Leo”). Participants first heard a note and were then presented with four buttons on the screen, with each button representing one of the four names. They were asked to respond by clicking on the button of the person who played the note. Participants were not told *a priori* which names were assigned to which notes and thus had to learn the note-name associations through feedback, which was provided after each response in audiovisual form (i.e., hearing the note again with visual feedback notifying participants which answer was correct). However, the associations of names and notes never changed throughout the experiment. The control condition had three levels of increasing difficulty, conceptually similar to the videogame condition. Figure 2 provides an overview of the progression of levels for the control condition.

Thus, although both conditions necessarily differed in their procedures, both groups completed a conceptually similar ‘rote test’ and ‘generalisation test’ which did not provide feedback, used the same timbres, and consisted of the same number of trials. As such, despite these differences in procedure across condition, the rote and generalisation tests were designed to facilitate comparisons across conditions, and as such performance on these components (compared to training, which differed substantially across conditions) is the focus of the present thesis. Additionally, it should be noted that both groups completed the same number of training and testing trials (168 in total).

Figure 2

Overview of the procedure for the control condition



Control Model

Training Phase

Hear a note, make a 4-alternative forced choice response as to who played that note (Sarah, David, Karen, Leo)

Receive auditory and visual feedback (correct button is circled and the note replays)

104 trials

Testing Phase

Rote Test

No feedback provided

Same notes as those heard in training

32 notes

Generalisation Test

No feedback provided

Sound ambiguous with respect to pitch height (Shepard tones) and never encountered in the training context

32 trials

Questionnaire and Debriefing

All participants completed a final questionnaire, administered after the AP training and testing paradigm. The questionnaire was composed of basic demographic questions, as well as questions meant to ascertain musical experience. Specifically, the questionnaire asked: age in years, gender, highest level of education, any training on musical instrument including voice (if "yes": main instrument, number of years of instruction, Age at which participants began instruction), possession of absolute/perfect pitch (yes, no, not sure), native language (English or other; if other, what language?). Following the questionnaire, participants were presented with a debriefing form, detailing the purpose of the study, and providing references for further reading.

At this time, participants were also given a unique completion code, which they entered into Mechanical Turk to confirm participation and to receive their payment.

Results

Assessing Accuracy against Chance Performance

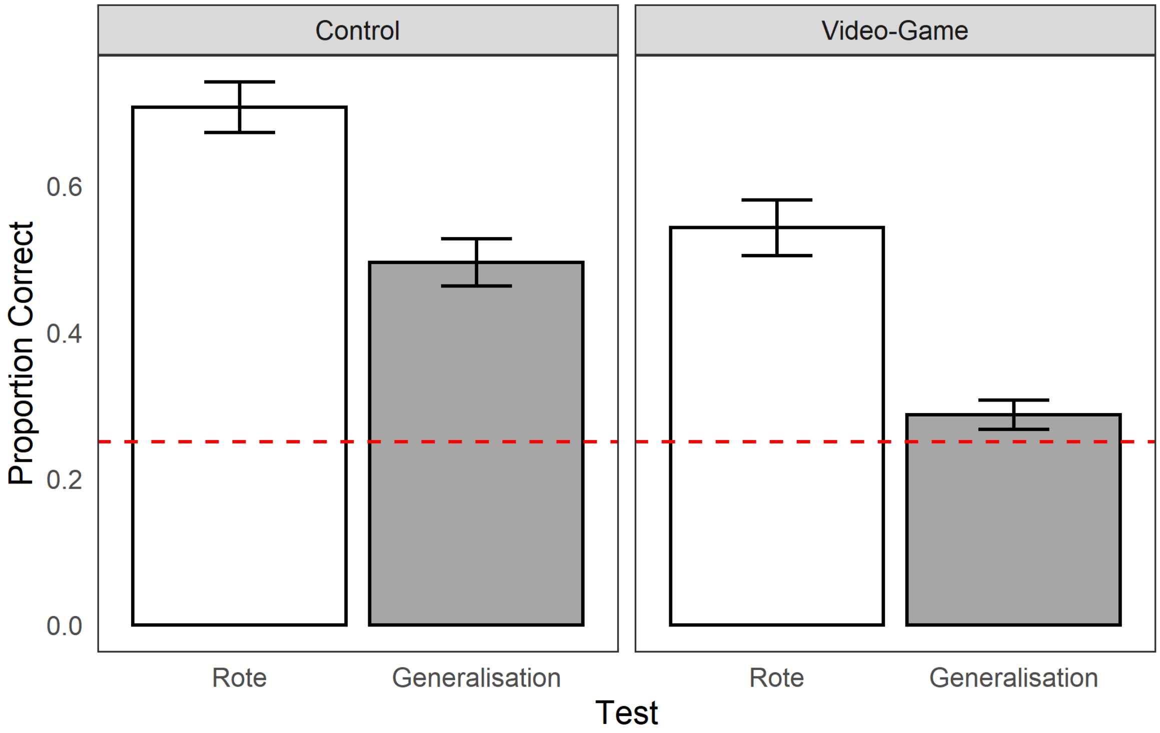
One-tailed, one sample t-tests against a known mean (25%) were used to assess whether performance significantly differed from chance. Although both conditions displayed significantly above chance performance for both the rote and generalisation tests, the effect sizes differed across both condition and test. Participants in the videogame condition achieved 54.28% in the rote test, which was significantly above the chance estimate and had a large effect size, *t*(45) = 7.72, *p* < .001, d = 1.14. For the generalisation test, performance for the videogame participants dropped to 28.74%, which was statistically significant but had a small effect size, *t*(45) = 1.87, *p* = .033, d = 0.28. Participants in the control condition achieved 70.72% in the rote test, which was significantly above the chance estimate and had an extremely large effect size, *t*(45) = 13.21, *p* < .001, d = 1.95. Although performance was attenuated in the generalisation test relative to the rote test for the control condition participants, overall performance was 49.52%, which was robustly above chance and had a comparably large effect size as the rote test performance in the videogame condition, *t*(45) = 7.64, *p* < .001, d = 1.13.

Assessing Accuracy across Condition and Test

Having established that all conditions and tests were independently above chance, performance differences across condition and test were formally assessed through a 2 (condition: videogame, control) x 2 (test: rote, generalisation) mixed ANOVA. The results of this ANOVA revealed a significant main effect of condition *F*(1,90) = 24.84, *p* < .001, with participants in the control condition (*M* = 60.1%, 95% Confidence Interval of 54.9% to 65.4%) outperforming participants in the videogame condition (*M* = 41.5%, 95% Confidence Interval of 36.3% to 46.8%). The ANOVA additionally revealed a significant main effect of test, with significantly higher performance on the rote test (*M* = 62.5%, 95% Confidence Interval of 58.0% to 67.0%) compared to the generalisation test (*M* = 39.1%, 95% Confidence Interval of 34.7% to 43.6%). There was no significant interaction of condition and test *F*(1,90) = 0.74, *p* = .391. The factors from the 2 x 2 ANOVA are plotted in Figure 3.

Figure 3

Accuracy plotted as a function of test (rote, generalisation) and condition (videogame, control)



Note: Error bars represent plus or minus one standard error of the mean. The dashed red line represents chance performance in terms of proportion correct (.25).

Self-Reported Engagement

Participants' self-reported engagement responses are summarised in Table 1. Although self-reported engagement was high across each measurement (attention to the task, willingness to continue, enjoyment of the task, self-assessed learning efficacy, and motivation to perform at one's best), as evidenced by mean ratings that were generally above the median point of the scale (i.e., 3), the only notable difference in engagement ratings across condition was the motivation measurement, with participants in the control condition reporting that they were marginally more motivated to perform well.

Table 1

Descriptive and inferential statistics comparing responses on the engagement questionnaire across conditions.

Prompt	**Videogame**	**Control**	***t*-value**	***p-value***
How effectively did this experience capture your attention?	4.48 (0.75)	4.61 (0.65)	0.89	.376
If given the opportunity, how likely would you be to continue engaging with this experience outside of this study?	2.98 (1.42)	3.11 (1.39)	0.44	.657
How much did you enjoy the experience?	3.85 (1.01)	3.83 (1.04)	-0.10	.919
How well do you think you learned as result of this experience?	3.43 (1.11)	3.24 (0.99)	-0.89	.375
How motivated were you to perform well?	4.41 (0.91)	4.72 (0.54)	1.95	.054

Note: Values in parentheses represent standard deviations.

Correlating AP Learning with Additional Variables

Exploratory analyses were conducted to determine whether AP learning in both the rote test and the generalisation test was correlated with the measured musical, language, demographic, and task engagement variables. Specifically, associations between AP learning and (1) age, (2) gender (woman, non-woman), (3) education (college degree or above, less than a college degree), (4) native English speaker (yes, no), (5) musical training (in years), and (6) responses from the engagement questionnaire were assessed. Separate analyses were conducted for each condition, given the significant performance differences observed across the videogame and control conditions.

Performance on the rote test was significantly correlated with performance on the generalisation test for both the videogame, *r*(44) = .37, *p* = .011, and control, *r*(44) = .43, *p* = .003, conditions, which is not surprising considering that both tests can be considered related subtests. For both the videogame, *r*(44) = -.30, *p* = .043, and control, *r*(44) = -.30, *p* = .044, conditions, there was an unexpected correlation between gender and rote test performance, with women showing lower performance than other participants. For the control condition, there were significant correlations between self-reported willingness to continue training and both the rote test, *r*(44) = -.34, *p* = .021, and generalisation test, *r*(44) = -.33, *p* = .025, with a *greater* willingness to continue training being associated with *worse* performance. No other variables were significantly associated with AP learning.

Discussion

The results obtained in this study were surprising for two main reasons. Not only did the participants in the control condition outperform the participants in the videogame condition (contrary to the main hypothesis), but the control condition participants also achieved some of the most impressive AP learning rates to date, given the relatively short timeframe of learning. Within approximately 50 minutes of learning, participants in the control condition were able to score an average of 70.72% correct answers on the rote test and an average of 49.52% correct answers on the generalisation test (with chance performance being 25%). While performances on pitch categorisation tasks in non-AP participants have been promising in past studies, pitch generalisation task performance in non-AP participants have historically been very poor, which might suggest that participants are learning other cues (e.g., pitch height) rather than pitch chroma (Bongiovanni et al., 2023).

In this context, the levels of performance seen in this study are unprecedented given the timeframe of learning. The ability to categorise Shepard tones is a good indicator of AP learning since the pitch height is ambiguous, meaning participants must use pitch chroma to accurately categorise the sounds. This can be visualised by looking at the performance of the videogame condition on the rote test. This study tested performance on four notes, meaning chance performance was 25%. However, performance at 50% on the rote test might have reflected an over-reliance on pitch height as a cue. Since in the videogame trial the four gates were lined up vertically (with the highest pitch at the top and the lowest pitch at the bottom) and the average rote test performance was 54.28%, it is possible that participants were attending to pitch height more so than participants in the control condition, which would potentially explain the near-change performance on the generalisation test (when pitch height cues were made to be unreliable). However, we did not perform a by-note analysis, nor did we ask participants to reflect on their learning strategies, so this remains speculative. Regardless, the 49.52% on the generalisation test in the control condition is even more impressive in this context.

The discrepancy between the results of the two conditions is rather puzzling. Why would the standard training paradigm perform better than the videogame paradigm, especially given prior research suggesting that gamification can facilitate perceptual categorisation (e.g., Lim & Holt, 2011)? Although both paradigms were well matched in their final rote and generalisation tests, one potentially major aspect in which the two paradigms differed was in difficulty during initial training. These initial differences in difficulty may have led to differences in psychological state, such as *flow* (Csikszentmihalyi, 1990), which in turn may have contributed to the unexpected learning differences between the videogame and control conditions.

The concept of flow is “a state of mind characterized by focused concentration and elevated enjoyment during intrinsically interesting activities” caused when integrating work and play (Hamari et al., 2016, p. 2). Flow is said to be an optimal state for learning, as participants are the most receptive to absorbing new information. A flow state can also be enhanced by certain properties of a task; one such property is the difficulty of the task. Should a task be too easy, the individual will lose interest and fail to learn from the experience. However, should the task be too difficult, the individual will become frustrated and incapable of learning. Within this framework, one possible explanation of the relatively poor learning from the videogame condition is that the paradigm was too easy in its initial training levels relative to its final testing levels. Indeed, a significant portion of the training trials (72 of 104) allowed participants to respond simply based on visual cues (i.e., the “lighting up” of the correct gates). Although the colour of the gates dimmed in the second level of training, they were still visible throughout the level, meaning participants in the videogame condition had relatively little experience making responses that were solely based on the trained sounds prior to the rote and generalisation tests.

In contrast, participants in the control condition had no initial means of knowing which button was correct for any given sound, as the associations were entirely arbitrary. While this might appear to challenge the framework of a flow state, as such a learning environment might be considered too difficult, one potentially critical factor was the fact that the present experiment only tested four (of the possible 12) Western note categories. In contrast, prior AP training research has trained all 12 note categories at once (e.g., Van Hedger et al., 2015; 2019) or trained six note categories at a time (e.g., Gervain et al., 2013), which might represent too many categories to initially test (i.e., representing training conditions that are too difficult to achieve a flow state). In contrast, other paradigms have focused on training a single note (e.g., Bongiovanni et al., 2023), which leads to excellent rote learning but near-chance generalization, potentially because the initial training environment is too easy. Thus, the present control condition may have inadvertently created optimal difficulty for learning and generalisation.

Another aspect of gamification that videogame developers have been using for decades and which was present in both conditions in the present experiment is a carefully calibrated increase in complexity. Currently, the sounds that participants were subjected to increase in complexity from the rote test to the generalisation test. However, the controls and mechanics of the model remained the same—they were still either moving a spaceship or clicking on a button labeled with a name. This presumably allowed participants to scaffold their initial experiences to then perform well in a new, more challenging environment (the generalization test), although participants in the videogame condition may have not been as effective in this scaffolding as their initial learning experiences may have been too easy and emphasized responding to the wrong cues (visual rather than auditory), as previously discussed. Nevertheless, considering both optimal difficulty and increases in complexity represents key factors to continue investigating in future work in this area. In support of this idea, some of the recent claims that AP is trainable in adulthood have used calibrated increases in complexity as a part of their training paradigms (Van Hedger et al., 2019; Wong, Lui et al., 2020; Wong, Ngan et al., 2020).

With all of this in mind, the present findings provide several promising avenues for future AP learning research to pursue. What is clear is that despite the advantages that the control condition had over the videogame condition, participants in the videogame condition still showed some AP learning, even in the generalisation test. Using what we have learned from this study, we believe we can improve the videogame model to at least match (and perhaps exceed) the results of the control condition. The most obvious change would be to increase the relative difficulty of the videogame condition: as discussed previously, there are several reasons to believe that the current version of the videogame paradigm was too easy, even to the point of hindering learning. While there are many ways of increasing difficulty, priority should be given to trying to emulate the same kind of difficulty present in the control condition. As stated above, the control condition was not told which name belonged to which note and needed to identify that on their own through trial and error. This is in stark contrast to the videogame condition, in which the associations between the gates and their respective notes were denoted by the colour of the gates and told to the participant explicitly via which gate was lit up. Therefore, in an effort to create a similar kind of difficulty, removing the explicitness of the note-gate association is advised. In practice, however, it is not immediately clear how this should be implemented. Although there are many ways to go about this (partly the reason why game design is so intimidating), one possibility would be to begin participants at level 3 of the training phase—removing the colours of the gates and providing visual and auditory feedback on whether they passed through the correct gate. Another possibility would be to keep the gate colours and instead increase the speed at which the spaceship travels, giving participants less time to decide on which gate to pass through, which is more akin to the approach taken by Lim and Holt (2011). However, this idea has the possibility of increasing reliance on the colours and decreasing their rate of AP learning, as the colours are easier to use as an indicator compared to sound. Speed could still be used as an intrinsic motivator: perhaps the colours of the gates are removed, but every time you pass through the correct gate the speed of the ship increases. To make this optimally challenging, the speed of the ship would also slow down every time the ship passes through the incorrect gate, essentially borrowing from adaptive staircase procedures used in psychophysics to ensure that the task is appropriately challenging for everyone (e.g., Leek, 2001). A similar approach has been recently implemented in AP training with encouraging learning trajectories as a result (Wong, Lui et al., 2020).

Whether these changes would improve AP learning, as opposed to just improving participant engagement, is still up for debate. Although there is an intuitive reason to predict that increased task engagement should lead to increased learning, the results of the present experiment did not support this (as participants in the control condition, who performed significantly better than participants in the videogame condition, did not generally self-report greater engagement with the task). It is possible that, in the present context, participant responses were influenced by demand characteristics (McCambridge et al., 2012) or even cognitive dissonance (e.g., Festinger, 1957), and thus may have not been sensitive to detecting potential engagement differences across conditions. In support of this possibility, the responses were very positive (average responses ranged from 2.98 - 4.72 on a 5-point scale). The most puzzling result of the engagement questions is the borderline significance of the responses to the question "How motivated were you to perform well?" in favour of the control condition. According to this result, the participants in the control condition—the hypothesized less appealing training paradigm—were slightly more motivated than the participants in the videogame condition—the hypothesized more appealing training paradigm. This result, although unexpected, fits within the other main unexpected finding of the experiment – that participants in the control condition outperformed participants in the videogame condition. Specifically, it is possible that due to the differences in difficulty across conditions, participants in the control condition had a greater investment to perform well in the task as their initial feedback would have suggested that the task was not easy and thus performing well would take work.

An important point in contextualizing the unexpected results from the control condition is to consider the relative magnitude of learning in relation to the study upon which the control condition was based (Gervain et al., 2013). The results of Gervain et al. (2013) were used to provide "proof-of-concept" support for a critical period in AP acquisition, through administering a drug (valproate) thought to re-open critical periods and finding that participants on valproate performed better at AP learning compared to when they were taking a placebo. In the study's own words, valproate is "a commonly used anticonvulsant and mood stabilizer, known to inhibit HDAC and modulate the epigenome to promote neuroplasticity" (Gervain et al., 2013, p. 2). Their study was designed with two within-participant conditions, a treatment condition that was administered valproate and a placebo condition (with the drug ordering counterbalanced). Importantly, they found that the participants taking valproate demonstrated better AP learning than when participants were taking a placebo. Gervain et al. (2013) also claimed to have trained genuine AP—a claim that was legitimised due to the pharmacological nature of the study. Although the exact magnitude of learning is difficult to compare, considering Gervain et al. (2013) trained six notes and the present experiment trained four, the present findings from the control condition (approximately 25 points above chance) was larger than the learning found in Gervain et al. (approximately 11 points above chance). Thus, despite testing adults who were not administered a drug to re-open a critical period of learning, the amount of AP learning in both our study and Gervain et al. (2013) is comparable. We believe this alone should cast doubt on the *necessity* of using a pharmacological intervention to train AP in an adult population. Another critical difference to note is that the Gervain et al. study lasted multiple weeks, with participants training for seven days (approximately 10 minutes per day), whereas our study lasted only 45 to 60 minutes and only training for around 30 to 45 minutes—again with comparable results. We would also suggest a reevaluation of the claim that participants in Gervain et al. (2013) learned AP (even as proof-of-concept), as it is much more accurately described as AP-like pitch categorisation.

The present study can therefore provide interesting introspection on the current state of AP discourse as well as research as a whole. Had we not conceptually replicated Gervain et al.'s (2013) training paradigm, we would not have been in a position to comment on the relative (in)effectiveness of the current videogame paradigm, nor would we have been able to challenge the assumption that pharmacological interventions are needed for successful AP learning in adults. We will also take this opportunity to assert that our study does not prove the trainability of AP, given that we only assessed learning on four (of 12) note categories and given that we did not assess learning stability over time, which is an important factor in determining genuine AP ability (e.g., Van Hedger et al., 2015). However, the present findings can be integrated into a larger body of research, which has emerged over the past several years, suggesting that AP may be trainable even in adulthood (Van Hedger et al., 2018; Wong, Lui et al., 2020; Wong, Ngan et al., 2020). These relatively recent studies, however, all implemented extensive training paradigms, lasting several hours each. As such, future research would benefit from examining exactly how elements of gamification (e.g., optimized engagement and difficulty that is calibrated to facilitate a flow state) can accelerate learning trajectories of AP. Despite the inconclusive results with respect to gamification, the present study emphasizes the need to continue reconsidering the upper limits of AP trainability into adulthood. AP especially was seen as an open-and-shut case with respect to adult acquisition; however, we believe this study provides sufficient evidence to reopen the investigation.

References

Baharloo, S. (2001). *Genetics of absolute pitch* [Ph.D., University of California, San Francisco]. https://www.proquest.com/docview/252316420/abstract/208C55A0A3F4456PQ/1

Cheng, M.-T., She, H.-C., & Annetta, L. A. (2015). Game immersion experience: Its hierarchical structure and impact on game-based science learning. *Journal of Computer Assisted Learning*, *31*(3), 232–253. https://doi.org/10.1111/jcal.12066

Csikszentmihalyi, M., & Csikszentmihalyi, I. S. (Eds.). (1992). *Optimal experience: Psychological studies of flow in consciousness.* Cambridge university press.

Cuddy, L. L. (1968). Practice effects in the absolute judgment of pitch. *The Journal of the Acoustical Society of America*, *43*(5), 1069–1076. https://doi.org/10.1121/1.1910941

Deutsch, D. (2002). The Puzzle of Absolute Pitch. *Current Directions in Psychological Science*, *11*(6), 200–204. https://doi.org/10.1111/1467-8721.00200

Dooley, K., & Deutsch, D. (2010). Absolute pitch correlates with high performance on musical dictation. *The Journal of the Acoustical Society of America*, *128*(2), 890–893. https://doi.org/10.1121/1.3458848

Dosher, B. A., & Lu, Z.-L. (1998). Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. *Proceedings of the National Academy of Sciences*, *95*(23), 13988–13993. https://doi.org/10.1073/pnas.95.23.13988

Gervain, J., Vines, B. W., Chen, L. M., Seo, R. J., Hensch, T. K., Werker, J. F., & Young, A. H. (2013). Valproate reopens critical-period learning of absolute pitch. *Frontiers in Systems Neuroscience*, *7*, 102. https://doi.org/10.3389/fnsys.2013.00102

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press.

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does Gamification Work? -- A Literature Review of Empirical Studies on Gamification. *2014 47th Hawaii International Conference on System Sciences*, 3025–3034. https://doi.org/10.1109/HICSS.2014.377

Hamari, J., Shernoff, D. J., Rowe, E., Coller, B., Asbell-Clarke, J., & Edwards, T. (2016). Challenging games help students learn: An empirical study on engagement, flow and immersion in game-based learning. *Computers in Human Behavior*, *54*, 170–179. https://doi.org/10.1016/j.chb.2015.07.045

Hartman, E. B. (1954). The influence of practice and pitch-distance between tones on the absolute identification of pitch. *The American Journal of Psychology*, *67*(1), 1–14.

Kuusi, T. (2009). *Tune Recognition from Melody, Rhythm and Harmony*.

Laamarti, F., Eid, M., & El Saddik, A. (2014). An Overview of Serious Games. *International Journal of Computer Games Technology*, *2014*, 1–15. https://doi.org/10.1155/2014/358152

Leek, M. (2001). Adaptive procedures in psychophysical research. *Perception & Psychophysics, 63*(8), 1279-1292. https://doi.org/10.3758/BF03194543

Li, J., & Gu, Z. (2021). Orpheus: A Voice-Controlled Game to Train Pitch Matching. In X. Fang (Ed.), *HCI in Games: Serious and Immersive Games* (pp. 33–41). Springer International Publishing. https://doi.org/10.1007/978-3-030-77414-1_3

Lim, S., & Holt, L. L. (2011). Learning Foreign Sounds in an Alien World: Videogame Training Improves Non-Native Speech Categorization. *Cognitive Science*, *35*(7), 1390–1405. https://doi.org/10.1111/j.1551-6709.2011.01192.x

McCambridge, J., de Bruin, M., & Witton, J. (2012). The effects of demand characteristics on research participant behaviours in non-laboratory settings: a systematic review. *PLOS One*, *7*(6), e39116. https://doi.org/10.1371/journal.pone.0039116

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2021). An online headphone screening test based on dichotic pitch. *Behavior Research Methods*, *53*(4), 1551–1562. https://doi.org/10.3758/s13428-020-01514-0

Miyazaki, K. (1993). Absolute Pitch as an Inability: Identification of Musical Intervals in a Tonal Context. *Music Perception*, *11*(1), 55–71. https://doi.org/10.2307/40285599

Mull, H. K. (1925). The Acquisition of Absolute Pitch. *The American Journal of Psychology*, *36*(4), 469–493. https://doi.org/10.2307/1413906

Nacke, L. E., & Deterding, S. (2017). The maturing of gamification research. *Computers in Human Behavior*, *71*, 450–454. https://doi.org/10.1016/j.chb.2016.11.062

Paney, A. S., & Kay, A. C. (2015). Developing Singing in Third-Grade Music Classrooms: The Effect of a Concurrent-Feedback Computer Game on Pitch-Matching Skills. *Update: Applications of Research in Music Education*, *34*(1), 42–49. https://doi.org/10.1177/8755123314548047

Peretz, I., Champod, A., & Hyde, K. (2003). Varieties of Musical Disorders: The Montreal Battery of Evaluation of Amusia. *Annals of the New York Academy of Sciences*, *999*, 58–75. https://doi.org/10.1196/annals.1284.006

Russo, F. A., Windell, D. L., & Cuddy, L. L. (2003). Learning the "Special Note": Evidence for a Critical Period for Absolute Pitch Acquisition. *Music Perception: An Interdisciplinary Journal*, *21*(1), 119–127. https://doi.org/10.1525/mp.2003.21.1.119

Takeuchi, A. H., & Hulse, S. H. (1993). Absolute Pitch. *The American Psychological Association*, *113*(2), 345–361.

Theusch, E. (2010). *The genetic epidemiology of absolute pitch* [Ph.D., University of California, San Francisco]. http://www.proquest.com/docview/849717001/abstract/BF71D15927A74B12PQ/1

Van Hedger, S. C., Heald, S. L. M., Koch, R., & Nusbaum, H. C. (2015). Auditory working memory predicts individual differences in absolute pitch learning. *Cognition*, *140*, 95–110. https://doi.org/10.1016/j.cognition.2015.03.012

Van Hedger, S. C., Heald, S. L. M., & Nusbaum, H. C. (2019). Absolute pitch can be learned by some adults. *PLOS ONE*, *14*(9), e0223047. https://doi.org/10.1371/journal.pone.0223047

Wade, T., & Holt, L. L. (2005). Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task. *The Journal of the Acoustical Society of America*, *118*(4), 2618–2633. https://doi.org/10.1121/1.2011156

Wayland, R., Herrera, E., & Kaan, E. (2010). Effects of musical experience and training on pitch contour perception. *Journal of Phonetics*, *38*(4), 654–662. https://doi.org/10.1016/j.wocn.2010.10.001

Wong, Y. K., Lui, K. F. H., Yip, K. H. M., & Wong, A. C.-N. (2020). Is it impossible to acquire absolute pitch in adulthood? *Attention, Perception, & Psychophysics*, *82*(3), 1407–1430. https://doi.org/10.3758/s13414-019-01869-3

Wong, Y. K., Ngan, V. S., Cheung, L. Y., & Wong, A. C. (2020). Absolute pitch learning in adults speaking non-tonal languages. Quarterly journal of experimental psychology (2006), 73(11), 1908–1920. https://doi.org/10.1177/1747021820935776

Yang, J.-S., & Cheng, C.-W. (2020). A Pitch Perception Training Game in VR Environment for Enhancing Music Learning Motivation. *2020 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*, 1–2. https://doi.org/10.1109/ICCE-Taiwan49838.2020.9257989

Curriculum Vitae

Name:	André Grenier
Place and Year of Birth:	Kitchener, Canada, 2001
Secondary School Diploma:	École Secondaire Monseigneur Bruyère, London, Canada (2019)
Presentations:	“Evidence for Absolute Pitch Learning Found in Non-Absolute Pitch Possessing Participants,” CURL Spring Conference. Huron University College, London, ON, April 2023