ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF ARTS AND SOCIAL SCIENCES

THE ACCEPTABLE PITCH RANGE(S) FOR SINGLE NOTE REPETITIONS IN MUSIC APTITUDE EXAMINATIONS

M.A. THESIS

Oğul KÖKER

Department of Music

Music Program

DECEMBER 2017

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL OF ARTS AND SOCIAL SCIENCES

THE ACCEPTABLE PITCH RANGE(S) FOR SINGLE NOTE REPETITIONS IN MUSIC APTITUDE EXAMINATIONS

M.A. THESIS

Oğul KÖKER
(409141112)

Department of Music

Music Program

Thesis Advisor: Assoc. Prof. Dr. Ozan Baysal

DECEMBER 2017

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ SOSYAL BİLİMLER ENSTİTÜSÜ

MÜZİK YETENEK SINAVLARINDA TEK SES TEKRARLARI İÇİN KABUL EDİLEBİLİR PERDE ARALIĞI (ARALIKLARI)

YÜKSEK LİSANS TEZİ

Oğul KÖKER
(409141112)

Müzik Anabilim Dalı

Müzik Programı

Tez Danışmanı: Doç. Dr. Ozan Baysal

ARALIK 2017

Oğul KÖKER (409141112), an M.A. student of ITU Graduate School of Arts and Social Sciences successfully defended the thesis entitled “THE ACCEPTABLE PITCH RANGE(S) FOR SINGLE NOTE REPETITIONS IN MUSIC APTITUDE EXAMINATIONS”, which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor : **Assoc. Prof. Dr. Ozan Baysal** ....
İstanbul Teknik Üniversitesi

Jury Members : **Assoc. Prof. Dr. Jerfi Aji** ....
İstanbul Teknik Üniversitesi

Assist. Prof. Dr. Deniz Tunçer ....
İstanbul Medeniyet Üniversitesi

Date of Submission : 17 November 2017
Date of Defense : 12 December 2017

FOREWORD

As one of the bursars of the TÜBİTAK (Türkiye Bilimsel ve Teknolojik Araştırma Kurumu [The Scientific and Technological Research Council of Turkey]) funded project titled "Müzik Algısı Ölçme-Değerlendirme Sınavlarında Ses-İşleme Araçlarının Kullanılabilirliği" [The Applicability of Sound-Processing Tools in the Exams for Music Perception Assessment and Evaluation] (Project no.: 215K017), I've been part of a research group of engineers and musicologists (namely, Prof. Dr. Turan Sağer, Prof. Dr. Nilgün Doğrusöz Dişiaçık, Assoc. Prof. Dr. Barış Bozkurt, Assoc. Prof. Dr. Ozan Baysal, and Burçin Bahadır Güner) since June 2016 with a short break of three months.

Focusing on the first phase of the entrance exams of İstanbul Teknik Üniversitesi Türk Musikisi Devlet Konservatuvarı (İTÜ TMDK) [Istanbul Technical University Turkish Music State Conservatory], this research group has been working tirelessly to improve the quality of such exams by assessing if it's possible to produce computer-based aids for that end. Briefly stated, this is done by combining computer-based analyses of the sound recordings taken at the entrance exams with the statistical information from the assessment forms filled by the jury members at those exams.

This thesis grew out from the questions that emerged from the research done as part of the project. Employing the sounds recorded by the research group, a survey was devised to find out the acceptable frequency ranges for the answers given to the single note repetition questions at the conservatory entrance exams.

Some parts of this thesis were done at the room which was assigned by the İTÜ TMDK Müzikoloji Bölümü [Department of Musicology] to be used for the aforementioned project.

I gratefully acknowledge the financial support of TÜBİTAK and institutional support of İTÜ TMDK Department of Musicology for this thesis. I'm grateful for the survey software written by Assoc. Prof. Dr. Barış Bozkurt, which saved me from doing tens of hours of laborious work. I express my sincerest thanks to those who participated in the survey with extreme patience and devotion. I'm indebted to my wife, Merve Doğan Köker, who tirelessly shared her knowledge on the subject with me. Finally, I thank my advisor Assoc. Prof. Dr. Ozan Baysal and all others worked as part of the research group – without them, I could not even imagine producing such a work.

December 2017 Oğul Köker

(Musicologist)

TABLE OF CONTENTS

Page

ABBREVATIONS

ADSR	**:** Attack, decay, sustain, release
JND	**:** Just noticeable difference
LSD	**:** Least significant difference
SPL	**:** Sound pressure level

SYMBOLS

F0 **:** Fundamental frequency
Hz **:** Hertz

LIST OF TABLES

Page

LIST OF FIGURES

Page

THE ACCEPTABLE PITCH RANGE(S) FOR SINGLE NOTE REPETITIONS IN MUSIC APTITUDE EXAMINATIONS

SUMMARY

This thesis grew out from the research questions that emerged from the research done as part of the project “Müzik Algısı Ölçme-Değerlendirme Sınavlarında Ses-İşleme Araçlarının Kullanılabilirliği” [The Applicability of Sound-Processing Tools in the Exams for Music Perception Assessment and Evaluation] funded by TÜBİTAK (Türkiye Bilimsel ve Teknolojik Araştırma Kurumu [The Scientific and Technological Research Council of Turkey]) (Project no.: 215K017).

Employing the sounds recorded for the aforementioned project, a computer-based survey was devised to find out the acceptable frequency deviation ranges for the answers given to the single note repetition questions, in which the prospect students of the conservatory try to reproduce a reference pitch played on the piano by their voices.

The survey was conducted on faculty members teaching music-related topics at the university, and on people who will be eligible to become faculty members soon, with 35 participants in total.

Two major outcomes arose from the data gathered through the survey. Firstly, it was found out that there was indeed an acceptable frequency deviation range for a pitch sung by a prospect student to be considered a successful repetition of the reference pitch played on the piano: The survey participants taking the role of the jury members at conservatory entrance exams accepted performances with a quite big range of frequency deviations from the reference pitch as successful. This range which participants deemed acceptable as successful was approximately 75 cents, distributed unevenly around the absolute correct pitch. For the deviations lower than the absolute correct pitch, the success rates did not fall under the 60 % threshold until the deviations got approximately as big as 45 cents. For the deviations higher than the absolute correct pitch, the success rates did not fall under the 60% threshold until the deviations got approximately as big as 30 cents.

Secondly, it was found out that besides the fundamental frequency of the sung note, variables such as pitch envelope, timbre, and octave differences also affect the perceived pitch. This means that at the conservatory entrance exams the jury members’ evaluation of the prospect students might not be fair, since the effects of such variables on the evaluation process are unknown.

MÜZİK YETENEK SINAVLARINDA TEK SES TEKRARLARI İÇİN KABUL EDİLEBİLİR PERDE ARALIĞI (ARALIKLARI)

ÖZET

Bu tez, TÜBİTAK (Türkiye Bilimsel ve Teknolojik Araştırma Kurumu) destekli "Müzik Algısı Ölçme-Değerlendirme Sınavlarında Ses İşleme Araçlarının Kullanılabilirliği" adlı proje (proje no. 215K017) kapsamında yapılan araştırma sırasında ortaya çıkan araştırma sorularından hareketle şekillenmiştir.

Bahsi geçen proje için kaydedilen sesler kullanılarak, konservatuvar giriş sınavlarında aday öğrencilerin piyanoda çalınan bir referans sesi şarkılayarak tekrar etmelerinin istendiği tek ses tekrarı sorularına verilen yanıtlar için kabul edilebilir frekans sapma aralıklarının bulunması için bilgisayar tabanlı bir anket tasarlanmıştır.

Anket, üniversitelerde müzikle ilgili alanlarda ders veren akademik personel ve pek yakında öğretim görevlisi olmaya hak kazanacak kimseler üzerinde uygulanmıştır. Toplam katılımcı sayısı 35'tir.

Anket yoluyla toplanan veriden iki ana sonuç ortaya çıkmıştır. İlk olarak, aday öğrenci tarafından şarkılanan perdenin piyanoda çalınan referans perdenin başarılı bir tekrarı sayılabilmesi için gerçekten de kabul edilebilir bir frekans sapma aralığı olduğu görülmüştür: Konservatuvar giriş sınavı jüri üyesi rolünü üstlenen anket katılımcıları, referans perdeden oldukça büyük frekans sapmalarına sahip performansları başarılı kabul etmişlerdir. Katılımcıları başarılı kabul edilebilir bulduğu bu aralık, mutlak doğru perdenin etrafında eşitsiz bir dağılımla, yaklaşık olarak 75 cent'tir. Mutlak doğru perdeden pes sapmalar için başarı değerleri, sapmalar yaklaşık -45 cent gibi bir değer alıncaya dek %60 sınırının altına inmemiştir. Mutlak doğru perdeden tiz performanslar için başarı değerleri, sapmalar yaklaşık +30 cent gibi bir değer alıncaya dek %60 sınırının altına düşmemiştir.

İkinci olarak, performanslar değerlendirilirken, şarkılanan notanın temel frekansı dışında, perde zarfı, tını, oktav farkları gibi değişkenlerin de algılanan perdeyi etkiledikleri görülmüştür. Bundan, konservatuvar giriş sınavlarında jüri üyelerinin aday öğrencileri değerlendiriş şekillerinin adilane olmayabileceği anlamı çıkmaktadır, zira bu değişkenlerin değerlendirme süreçlerini nasıl etkilediği bilinmemektedir.

1. INTRODUCTION

This thesis grew out from the questions that emerged from the research done as part of the TÜBİTAK (Türkiye Bilimsel ve Teknolojik Araştırma Kurumu [The Scientific and Technological Research Council of Turkey]) funded project titled "Müzik Algısı Ölçme-Değerlendirme Sınavlarında Ses-İşleme Araçlarının Kullanılabilirliği" [The Applicability of Sound-Processing Tools in the Exams for Music Perception Assessment and Evaluation] (Project no.: 215K017)[1].

Employing the sound recordings done as part of the aforementioned project, a computer-based survey was devised to find out the acceptable frequency deviation ranges for the answers given to the single note repetition questions take part at conservatory entrance exams, in which the prospect students of the conservatory try to reproduce a reference pitch played on the piano by their voices.

To this end, the first phase of the entrance exams of İstanbul Teknik Üniversitesi Türk Musikisi Devlet Konservatuvarı (İTÜ TMDK) [Istanbul Technical University Turkish Music State Conservatory] was chosen as a case study[2].

---

[1] As a bursar of this project, the author of this text has been part of a research group of engineers and musicologists (namely, Prof. Dr. Turan Sağer, Prof. Dr. Nilgün Doğrusöz Dişiaçık, Assoc. Prof. Dr. Barış Bozkurt, Assoc. Prof. Dr. Ozan Baysal, and Burçin Bahadır Güner) since June 2016 with a short break in early 2017. The work of this group then can be summarized as combining computer-based analyses of the sound recordings taken at the entrance exams with the statistical information from the assessment forms filled by the jury members at those exams to see what kind of aids would be beneficial in the entrance exams and if it's possible to create such aids on the computer.

[2] The first phase of the entrance exams of İTÜ TMDK consists of several sections: (i) Pitch repetition, in which the prospect students try to reproduce a reference pitch played on the piano by their voices; (ii) dyad repetition, in which the prospect students try to reproduce a harmonic dyad played on the piano by their voices intervalically; (iii) triad repetition, in which the prospect students try to reproduce a harmonic triad played on the piano by their voices intervalically; (iv) tonal melody repetition, in which the prospect students try to reproduce by their voices a melody in major or minor tonalities played on the piano; (v) modal melody repetition, in which the prospect students try to reproduce by their voices a melody in modes reminiscent of Makams in Turkish music played on the piano; (vi) regular rhythmic repetition, in which the prospect students try to reproduce by hitting a table with a pen a divisive rhythm played by a jury member by hitting a table with a pen; (vii) irregular rhythmic repetition, in which the prospect students try to reproduce by hitting a table with a pen an additive rhythm played by a jury member by hitting a table with a pen.

When designing and conducting the survey, free software was employed as much as possible to make sure that if a researcher even wants to reproduce the survey, he/she should be able to do so with a small budget. Consequences of this choice will be discussed in the following chapters.

Following this introductory chapter comes Chapter 2, which contains a literature review, in which it is argued that there are no precedent studies similar to this thesis, but the literature on pitch perception might prove helpful to get a general understanding on the subject. Chapter 3 describes how the survey was designed and conducted. Chapter 4 offers the findings of the statistical analyses conducted on the data gathered via the survey. Chapter 5 discusses how these findings should be interpreted, and what these interpretations suggest for further research and for the improvement of conservatory entrance exams.

2. LITERATURE REVIEW

No other study dealing with the conservatory entrance exams that take place in Turkey was found. Nevertheless, there's a body of work regarding the entrance exams of music-related departments in fine arts faculties, and of the music education departments in faculties of education in Turkey. For example, there are studies that deal with standardization of entrance exams (Atak Yayla and Yayla, 2009), reliability of the entrance exams (Atar et al., 2013), and inter-jury reliability in the entrance exams (Ece and Kaplan, 2008).

Among these studies, however, there isn't one study that shares similarities with this thesis. The distinctness of this thesis arises from its focus on the sound, while all others focus on textual data.

Additionally, no other study dealing with the conservatory entrance exams in countries other than Turkey was found. Nevertheless, there's a body of work regarding the tests designed to evaluate musical skill, expertise, aptitude, and such[3].

Two recent tests somewhat comparable with conservatory entrance exams were mentioned in the relatively recent studies in Europe. One of these recently mentioned tests was the Musical Ear Test (Wallentin et al., 2010).

This test consists of 104 trials where participants were asked if the two consecutive phrase pairs they've heard are identical or not (Wallentin et al., 2010, p. 189). Trials consisting single pitches instead of phrases are non-existent.

Wallentin et al. (2010) state that even though the "correlation between musical expertise and music processing" is "well-documented", "many, if not most, cognitive investigations of music actually leave out individually measures of musical expertise".

---

[3] It should be noted that the literature regarding the older tests such as the famous one by Carl Seashore were overlooked in the context of this thesis, not because these tests are outdated, but because a historical account and assessment of such tests would be much beyond the scope of this thesis.

They argue that “the reason for this may be that an easily available, well-documented, objective, short, up-to-date test that reliably distinguishes musical expertise is lacking”. They state that the Musical Ear Test meets these criteria (p. 188).

They report that the three experiments done prove that this test “clearly distinguishes even small groups of professional-, amateur- and non-musicians; […] correlates with the imitation test, used both at the academies of music in Denmark at entrance, mid-term and final exams, and in several studies of musical expertise […]; and […] correlates with amount of musical practice (Wallentin et al., 2010, p. 188, 195).

According to Wallentin et al. (2010), the aforementioned imitation test, where the test participants get asked “reproduce short rhythmical and melodic phrases using hand-claps and singing”, is not fit for “testing musical perception in amateur- and non-musicians”, because not only it “often yield[s] a floor effect”, but also it may be not “optimal” for “experimental purposes”, “since it involves an element of subjective judgment of participants' imitation” (p. 189)[4].

The other of these recently mentioned tests was the Swedish Musical Discrimination Test (Ullén et al., 2014).

Ullén et al. (2014) state that the “endeavours to objectively measure musicality since the early 20th century” have been many. Yet, according to them, the “correlations between different tests and between tests and criteria such as teacher’s ratings and music school grades” are “relatively moderate”, because “different tests use different operationalisations of musicality”. In fact, according to them, there are two main approaches for testing “musicality” (p. 87).

“The atomistic approach”, say Ullén et al. (2014), “is based on the assumption that musicality is made up of several relatively narrow and distinct musical abilities”. According to them, “tests in this tradition have typically focused on basic sensory abilities, such as discrimination of various musically relevant sound stimuli” (p. 87).

---

[4] On one hand, one could argue that this test might be analogues to the conservatory entrance exams in Turkey, but since it does not contain a single note repetition section, we’ve ignored it in the context of this thesis.

Ullén et al. (2014) reports that although "empirical data" suggest that there's some truth to the "idea of independence of musically relevant perceptual abilities", "individual differences in discrimination tasks are […] also influenced by more general factors. "In fact", they state, "auditory discrimination tasks positively correlate with a broad range of non-musical cognitive tasks and psychometric modelling shows that general intelligence […] is an important factor underlying the positive covariation between different 'atomistic' tests of musical discrimination" (p. 87) [Single quotation marks are of original text].

"The omnibus approach", on the other hand, considers "musicality" as "a general high-level ability" (Ullén et al., 2014, p. 87). According to Ullén et al. (2014), "tests developed within this tradition are less concerned with characterizing components of musicality but rather tend to use a holistic approach where complex, acquired musical knowledge is assessed". "Typical […] items" in the tests based on this approach, say Ullén et al., "may involve quality judgments of musical performances or the production of musically meaningful responses to stimuli" (pp. 87-88).

The differences between the two approaches "mean that the omnibus tests typically are aimed at practicing musicians, while the atomist tests can be used for a wider range of purposes" (Ullén et al., 2014, p. 88).

Ullén et al. (2014) add that there is a third group of "musicality tests that do not easily fit into either of these main traditions as they focus on musical engagement, motivation and interests rather than the cognitive capacity to process musical information as such" (p. 88).

"The Swedish Musical Discrimination Test", with its purpose "to provide measures of basic aspects of musical ability operationalised as discrimination ability for auditory musical stimuli", employs the atomistic approach (Ullén et al., 2014, p. 88). Ullén et al. (2014) reports that in designing this test, they aimed to produce a test with a "short test-taking time, allows for online administration, and has a suitable difficulty level for general musically untrained populations in industrialized countries" (p. 88).

There were 6881 participants in this test conducted as part of a larger online survey (Ullén et al., 2014, p. 88). The pitch-related section of the test consisted of sine waves. In this section, participants listened pairs of tones; one with F0 of 500 Hertz (Hz), other with its F0 being in the range of 501-517 Hz. The order of tones in each pair were randomized. The task was to indicate if the first tone was higher or lower than the other tone with the difficulty getting “increased progressively by making the pitch difference between the notes smaller”. There were two other sections dealing with the domains of melody and rhythm (Ullén et al., 2014, p. 88).

Ullén et al. (2014) reports that the “internal consistencies and split-half reliabilities were excellent for all three scales” (melody, rhythm, and pitch). They state that all three scales “were associated with both musical experience, i.e. having played a musical instrument, as well as formal musical education” (p. 90).

Ullén et al. (2014) reports that “preliminary analyses showed that sex had a small mean effect on Pitch […] with a slightly lower mean for females […] than for males […]”, while “age showed a significant mean effect on Rhythm […] and Pitch […], with decreased discrimination skills with increased age” (p. 91). Thus, it seems Ullén et al. (2014) managed to archive their goal to produce a short online test that “has a suitable difficulty level for general populations” (pp. 88, 92).

It should be now clear that neither of these two tests share similarities in their structure or aim with the conservatory entrance exams in Turkey. The unique content and function of the conservatory entrance exams renders it harder to do any comparative study regarding its reliability, efficiency, and fairness.

Along with the literature regarding the tests, a review of literature regarding how pitch is perceived should also prove to be beneficial when interpreting the data gathered through the survey.

Oxemham (2012) states that pitch is one of the three “primary auditory sensations”, the other two being loudness and timbre (p. 13335). Similarly, Kollmeier et al. (2008) state that “the transformation of acoustical signals into auditory sensations can be characterized by psychophysical quantities such as loudness, tonality, or perceived pitch” (p. 61).

“Put it simply”, says Oxemham (2012), “pitch is the perceptual correlate of the periodicity […] of an acoustic waveform”. According to Oxemham, most common “pitch evoking sound” is “a harmonic complex tone” (p. 13335):

> This periodic waveform repeats at a rate corresponding to the fundamental frequency (F0) and can be decomposed into sinusoidal harmonics or overtones, which have frequencies at integer multiples of the F0 […]. The relative amplitudes of the harmonics within a complex tone play an important role in determining the sound quality, or timbre, of a sound. Despite differences in timbre and loudness, two tones generally have the same pitch if they share the same F0 (Oxemham, 2012, p. 13335).

Oxemham (2012) reports that "although young humans with normal hearing can hear sounds with frequencies between ~20 and 20,000 Hz, only repetition rates between ~30 and 4000 Hz elicit a pitch sensation that is salient enough to carry melodic information" (p. 13335).

Oxemham (2012) states that perceived pitch of "a tone" can retain even when "all the energy at the F0 is removed or masked by noise (p. 13335):

> From a perceptual standpoint, it makes sense that the pitch of a sound remains constant after the lowest harmonic components are removed or masked (occluded), so that some degree of perceptual invariance of a sound source can be maintained in a cluttered acoustic environment (Oxemham, 2012, p. 13335).

Oxemham (2012) states even though pitch is often considered as independent from other "perceptual dimensions, such as loudness and timbre", "some interactions" among them exist. He reports that "intensity" (loudness) and "brightness" (timbre) can affect the pitch perceived. In fact, detecting small pitch differences between sounds with different timbres was found to be hard even among the musically trained listeners (p. 13337). In this regard, Zarate et al. (2013) reports that recent studies found that (i) the manipulation of timbre (between dull and bright) affects the perception of the melodic interval size, (ii) both musicians and non-musicians discriminate intervals better when complex sounds are used instead of pure tones, and (iii) musicians and non-musicians have similar thresholds of pitch and interval discrimination for both pure and complex tones (p. 1). Similarly, according to Vurma (2014), recent studies show that timbral differences of the sounds with the same F0 value can be perceived as pitch-shifts by musicians and non-musicians (p. 1).

Vurma (2014) also reports that studies show that increasing SPL (sound pressure level) values of "pure tones" causes the perceived pitch to get higher for the F0 values above 4000 Hz and get lower for the F0 values below 2000 Hz. She adds that this effect is “less pronounced in the case of complex tones” (p. 2), so we can safely state that the effects of any SPL changes can be overlooked in this thesis, since our material is composed of complex sounds generated by piano and voice.

Kollmeier et al. (2008) report that when the frequency of a low frequency sine wave gets gradually increased, the perceived pitch increases linearly with that frequency up until 500 Hz threshold. When the frequency goes above 1000 Hz, however, the pitch perceived increases “approximately logarithmically with the increasing pitch” (p. 64).

According to Kollmeier et al. (2008), this relationship between the frequency and the perceived pitch resembles “the mapping of frequencies on the basilar membrane” inside the cochlea, where one half of the membrane reserved for the frequencies up to 2000 Hz, and the remaining frequency range of 2000-20000 Hz is handled by the other half. Similarly, JND (just noticeable difference) values for frequency for such sine wave is reported to be 3 Hz for frequencies below 500 Hz (thus, linear), and %0,6 for frequencies above 1000 Hz (hence, logarithmic). Kollmeier et al. add that as the frequency of sine waves increases, their pitch strength decreases steadily (pp. 64-65).

Kollmeier et al. (2008) state that compared to the perception process regarding sine waves, “a much more distinct pitch perception” occurs when one deals with broad band sounds such as musical instruments and “voiced speech elements”. According to them, the pitch perception for these complex sounds results at lower frequencies from temporal cues, and at high frequencies from spectral cues. They report, however, that “for the perception of the pitch frequency range of normal speech with F0s between approximately 80 Hz and 500 Hz, […] predominantly temporal cues are exploited by our ear”. For this range, they say, JND for pitch is approximately 1 Hz – meaning that the resolution for the F0s of complex sounds surpasses the resolution for the single sine waves (p. 65).

As part of an experiment, Zarate et al. (2013) created a sine wave with F0 of 349,23 Hz, corresponding to “an F4 in Western music”. Additionally, three instrument sounds (piano, flute, and voice) were generated using an electric keyboard at the same F0 value. Using these four sounds with their common F0 value, additional tones were created with F0 values deviating from the original F0 value with increments of 25 cents. All these tones eventually were paired up to create intervals between the range of 0 to 100 cents. Zarate et al. made a group of musicians and non-musicians (29 subjects in total) listen to 500 different pairs made out of these interval pairs and asked them which ones contained the bigger interval employing a “two-alternative forced-choice design” (pp. 2-3).

Pitch discrimination threshold was found to be 14,6 cents for musicians and 44,1 cents for non-musicians by Zarate et al. (2013). Musicians discriminated better all intervals compared to non-musicians. The pairs containing pure tones especially were discriminated by musicians much better than non-musicians. Among four timbres, the pairs with pure and piano tones were much more accurately discriminated by non-musicians interval-wise. Both musicians and non-musicians discriminated more accurately as the intervals got bigger (pp. 3-4).

Except for the intervals of 25 cents, the interval-discrimination was best with pure tones. Sensitivity to F0 changes was lowest for the flute tones. Accuracy for F0 changes was lowest for the voice tones (Zarate et al., 2013, p. 5).

A bias towards choosing the second interval of the pairs as the larger one was observed at all intervals, with non-musicians being more biased towards this decision than musicians. Musicians’ bias decreased as the intervals grew, but there was not much difference between the intervals of 50 cents and 75 cents. For non-musicians, bias decreased significantly only for the intervals of 100 cents (Zarate et al., 2013, p. 5).

For the intervals of 75 cents and above, the bias was lowest for the pure tone. For all intervals, the bias was lowest for the pure tone compared with flute and voice tones. When there was no difference between F0 values, compared to musicians, non-musicians tended to pick intervals containing instrument tones as the bigger intervals (Zarate et al., 2013, p. 5).

Thus, subjects did better and faster with pure tones compared with the complex instrumental tones (Zarate et al., 2013, p. 7). Zarate et al. (2013) argues that this unexpected result might have come into existence because they sequenced different timbres back to back hoping that those would get perceived as constituting intervals, but since such differences violated the expectancies of the participants, they might have caused distractions during the discrimination process, resulting a better performance with pure tones, even though they are known to be harder to interpret.

They add that this might have happened also because the relatively more frequent introduction of the pure tones compared to other tones during the experiment resulted in “a practice effect” (p. 7).

Zarate et al. (2013) report that “pitch changes seem to be best perceived (regardless of timbre) when F0 changes by at least 4%; the perception of smaller F0 changes is more influenced by timbre changes”. Yet they point out the fact that two of the intervals (namely, 25 cents and 50 cents from the 349,28 Hz) employed in their study were smaller than a difference of 4% from the original frequency, forcing the participants focus on timbre to detect pitch changes instead of focusing pitch itself (pp. 7-8).

As part of an experiment, Vurma (2014) made 31 professional musicians (16 pianists and 15 string players) to listen to 528 pairs of musical notes generated by digitally manipulated recordings of a viola. There were three distinct “pitch regions” (D#3, D4, and C#5) employed with their F0s manipulated in different ways (-60, -40, -20, 0, +20, +40, +60 cents). “Timbral succession” of the pairs were also altered: Bright-dull, dull-bright, bright-bright, dull-dull. Additionally, a silence of 3,5s was introduced in between some pairs (pp. 2-3).

Vurma (2014) reports that the participants tended to miss F0 differences when comparing identical timbres, while they gave false alarms for pairs with different timbres (p. 11).

The sensitivity for F0 differences was at its highest when both tones had bright timbres, and was at its lowest when only the tone with lower F0 had the bright timbre (Vurma, 2014). When the tone with higher F0 had the brighter tone, or both timbres were identical, participants detected the differences of 20 cents and more. When the tone with lower F0 had the bright timbre, however, they failed to detect the differences of 20 cents, and some of them even failed to detect the differences of 40 cents (p. 11).

Vurma (2014) reports that string players scored better in her study compared to pianists. She argues that the viola sounds employed in the test might have caused this difference (p. 11).

Vurma (2014) reports that false alarms increased when no silence was present in between the tones. She adds that the sensitivity towards differences was better when there was a 3.5s silence between tones. She states that this shows that "our nervous system requires some time in order better to separate pitch information from timbral information" (p. 11).

3. MATERIAL AND METHODOLOGY

3.1 Design of the Question Set

Reviewing the recordings done as part of the aforementioned project, the recordings (in 16-bit PCM wave format, stereo) of 6 different voices that didn't have much interfering outside noises such as people walking in the corridors or practicing their instruments in one of the practice rooms were picked. From these 6 recordings, 57 single-note recordings that had no trace of any outside noises were produced using the free audio editing software Audacity 2.1.2 (in 16-bit PCM wave format, mono).

Using the software Tony (Mauch et al., 2015) in its stock settings, the single-note recordings were analyzed to see their content with respect to pitch. The recordings containing more than one pitch according to Tony were eliminated from our recording pool (15 of them, to be exact), since such recordings would introduce random variables we could not control in our data.

From the remaining 42 recordings, 12 recordings (one half containing male, the other containing female voices) that contained pitches approximating the pitches of the equal temperament (Url-1) with the tolerance limit of 1 cent were picked. It should be noted that 12 recordings don't necessarily mean 12 different pitches. Instead of employing each pitch possible, examples of the same pitch class from different octaves were included to have data for future studies to see if the note register influences the evaluation of single note performance.

These 12 recordings would be the examples of successful performances to be presented in the survey. To be able to examine the range of tolerable F0 deviations for the single note repetitions when evaluating performers in the conservatory entrance exams, recordings of faulty performances with various controllable levels of pitch deviations were also needed. Creating such recordings by processing the recordings of successful performances already available was an easy choice, since it was thought that it would be easier than recording audio with such content.

For the faulty performances, going far than ±49 cents from the original pitch didn't make much sense, since a -/+50 cents difference stands at the center of two different pitches and any differences beyond a ±50 cents limit would mean that we were now in the domain of a different pitch. Since Zarate et al. (2013) reports that the pitch discrimination threshold was found to be 44,1 cents for even non-musicians (p.3), going beyond 45 cents would be an excess in this survey, in which every participant would have musical expertise. Thus, it was determined that the maximum deviation from the original pitch should be ±45 cents, leading to a total deviation range of 90 cents.

Zarate et al. (2013) also reports that the discrimination threshold was found to be 14,6 cents for musicians (p.3). Rounding up this threshold to 15 cents, it was decided that the range of 90 cents should be divided in a way that there is a spot for the deviations of 15 cents. The common divider of 5 was considered, but since it would lead to an abundance of material, it was decided that the range of 90 cents should be divided into steps of 7,5 cents, producing 12 different out of tune versions of the original pitches (Table 3.1).

Table 3.1 : Possible versions of each original recording; 13 in total.

(-) Deviations (in cents)						**Original Pitch**	**(+) Deviations (in cents)**					
-45	-37,5	-30	-22,5	-15	-7,5	No deviation.	+7,5	+15	+22,5	+30	+37,5	+45

With the help of the free online tool titled "Changing of the frequency about a cent value" (Url-1), the F0s for the out of tune versions to be created were calculated. Using the "Change Pitch…" effect in Audacity, 12 out of tune versions (in 16-bit PCM wave format, mono) were realized for each of the 12 original recordings. At this stage, 156 performance recordings (12 originals, along with 12 additional versions of each) to be employed in the survey were available.

Then two distinct problems occurred. Firstly, when looking over these 156 recordings on Tony to see if any human error was made, it was seen that Tony didn't offer a way to check if our 7,5 cent steps were implemented free of error, since it showed information regarding cents only in whole numbers, rounding the decimals to the closest whole number. Thus, the deviations of 22,5 cents, for example, would read either 22 cents or 23 cents.

Secondly, it was seen that Audacity's pitch shifting tool worked with much less accuracy than it was hoped. For example, when one of the successful recordings was picked and got processed to produce a version of it that's +45 cents higher, the rendered recording file read a couple of cents higher or lower on Tony. Eventually, it was decided that Audacity's pitch shifting capabilities lacked the quality needed for such a process, since no human error was found[5].

To solve these two problems, each deviation step (such as +7,5 cents, -22,5 cents, +30 cents, -45 cents and so on) was treated not as a spot based on a fixed deviation ratio, but the center of a region that spreads over a small range of cents. So, for example, a sound that was processed to be 30 cents lower than an equal temperament pitch, but turned out to be only 29 cents lower was considered to be part of a deviation level region that included every out of tune recording produced to be 30 cents lower, be it on spot or a bit off. Thus, based on the 12 different steps of deviation, 12 different regions of deviation were produced (Table 3.2).

[5] At this point, it could be possible to redo everything done so far using paid professional software to be able to produce the material with much more precision. Yet the idea of paid software being an obstacle for further research by preventing the reproduction of the setup of this study seemed also problematic, so it was decided that the usage of the free software should be continued.

Table 3.2 : Regions and their content.

(+) Deviation Regions	Aimed Deviation Amount (in cents)	Actual Deviation Amount (in cents)
1	+7,5	+6
		+7
		+8
2	+15	+14
		+15
3	+22,5	+21
		+22
		+23
4	+30	+29
		+30
		+32
5	+37,5	+37
6	+45	+44
		+45
		+46
		+47

(-) Deviation Regions	Aimed Deviation Amount (in cents)	Actual Deviation Amount (in cents)
1	-7,5	-7
		-8
2	-15	-14
		-15
3	-22,5	-20
		-23
4	-30	-28
		-29
		-30
5	-37,5	-37
		-38
6	-45	-45
		-46

Other Regions	Deviation Amount (in cents)
0	0
Control	>100

Using the free virtual studio technology called Iowa Piano in the free digital audio workstation titled Tracktion 5, eight piano recordings (in again 16-bit PCM wave format, mono) containing single piano notes were produced. The reason a different piano recording for each of the 12 original performance recordings wasn't produced is the fact that while listening to more than 1000 unique performances throughout the aforementioned project, it was never encountered a female singer who didn't reproduce what she heard on the piano exactly on the same octave, and it was never encountered a male singer who didn't reproduce what he heard on the piano an octave lower. Thus, in the pursuit of simulating an exam performance, a single pitch on piano could be both used as the reference for female singing the same pitch and as the reference for male singing that pitch one octave lower.

It should also be noted that this octave difference means that when talking about the deviation regions, the word pitch implies the pitch classes (such as C, E, G#, etc.) instead of the absolute pitches themselves (such as A4, middle C, F#3 etc.).

Since pianos do not get tuned in equal temperament in mathematical sense, the same processes that were employed to produce out of tune performance recordings were used on the 8 piano recordings to make them comply with equal temperament as much as the software allowed.

After that piano recordings were combined with performance recordings to produce fake exam performances, using Tracktion 5 again. Thus, 156 recordings that contain both successful and relatively unsuccessful single-note repetition performances were produced. As an addition to these recordings, 9 recordings that contain piano references followed by random pitches that constitute diatonic intervals with the reference pitch were produced. These intervals going well beyond 45 cents were obviously unsuccessful as single-pitch repetition performances. Thus, it was thought, any participant who would give these performances a passing grade, either lacked skill or focus. Whichever the case, however, such a person should not be part of a fair exam jury. Hence, the data gathered from him/her wouldn't be used in the context of this thesis.

All these recordings containing pairs of references and performances were looked over to see if there was any human error made during their preparation. No human error was found, but it was realized that when these files got rendered new unwanted deviations in the pitch were developed. It was thought that that this issue came into existence either because of the shortcomings of Tracktion 5's rendering tool, or simply because the sounds were too over-processed at this moment. Most of these new deviations were smaller than a cent and thus, they didn't affect the pace of our study. The ones which deviated more than 1 cent from the aimed pitches were excluded from the material to be used in the survey.

After this elimination, more than 100 recordings suitable for use were left in the question pool. Employing all of them would make the survey too long and thus add uncontrollable variables such as boredom to the data. So, it was decided to pick a certain number of performances to be included in the question set from each of the deviation regions. These questions were semi-randomized based on the following criteria: Disregarding octave differences, no two same notes should be presented back to back; no two recordings from the same region should be presented back to back; same gendered voices should not be heard more than 3 times back to back. Additionally, a group of questions were ordered in a way to create question groups that start with small deviations gradually getting bigger, and questions groups that start with big deviations gradually getting smaller, making sure that data on question ordering would be available if ever needed. Some other groups were also created disregarding the criteria above and going totally random. The control questions that contain deviations bigger than 100 cents were placed in certain spots to function as distractors. The last 13 questions of the survey were duplicates of some questions included earlier in the survey to make sure that a second line of control questions were available if necessary. Ultimately, the question set of 74 questions were created (Table 3.3).

Table 3.3 : Final version of the question set.

Question			Reference Piano Sound				Voice Performance					
No	Region	Additional Info	Pitch	Ideal Hz	Actual Hz	Difference between ideal and actual piano frequency (in cents)	Pitch	Aimed Deviation from the ideal piano reference (in cents)	Ideal Hz	Actual Hz	Actual deviation from the ideal piano reference (in cents)	Sex
1	Control	Perfect 4th	D#4	311,127	311,208	<1	G#4	+500	415,305	415,383	+500	Female
2	0	N/A	C5	523,251	523,262	<1	C5	0	523,251	523,285	<1	Female
3	+1	N/A	A#4	466,164	466,073	<1	A#3	+7,5	234,094	234,079	+7	Male
4	+2	N/A	F#4	369,994	369,907	<1	F#4	+15	373,214	373,201	+15	Female
5	+3	N/A	C4	261,626	261,557	<1	C3	+22,5	132,524	132,537	+23	Male
6	+4	N/A	A4	440,000	440,082	<1	A4	+30	447,691	447,630	+30	Female
7	+5	N/A	F#4	369,994	369,907	<1	F#3	+37,5	189,048	188,970	+37	Male
8	+6	N/A	A4	440,000	440,108	<1	A3	+45	225,793	225,714	+44	Male
9	Control	Minor 6th	C4	261,626	261,557	<1	G#4	+800	415,305	415,375	+800	Female
10	0	N/A	D#4	311,127	311,208	<1	D#4	0	311,127	311,076	<1	Female
11	-1	N/A	F#4	369,994	369,907	<1	F#3	-7,5	184,197	184,207	-7	Male
12	-2	N/A	C5	523,251	523,262	<1	C5	-15	518,737	519,000	-14	Female
13	-3	N/A	A4	440,000	440,108	<1	A3	-22,5	217,159	217,084	-23	Male
14	-4	N/A	F#4	369,994	369,907	<1	F#4	-30	363,638	363,781	-29	Female
15	-5	N/A	C4	261,626	261,557	<1	C3	-37,5	128,010	128,014	-37	Male
16	-6	N/A	D#4	311,127	311,208	<1	D#4	-45	303,144	303,208	-45	Female
17	+1	N/A	A4	440,000	440,082	<1	A4	+7,5	441,910	441,821	+7	Female
18	+4	N/A	A4	440,000	440,108	<1	A3	+30	223,846	223,735	+29	Male
19	+6	N/A	C4	261,626	261,557	<1	C3	+45	134,258	134,285	+45	Male
20	-5	N/A	F#4	369,994	369,907	<1	F#4	-37,5	362,066	362,090	-37	Female
21	Control	Minor 7th	C4	261,626	261,557	<1	A#4	+1000	466,164	466,428	+1001	Female
22	0	N/A	A4	440,000	440,082	<1	A4	0	440,000	440,083	<1	Female
23	+1	N/A	F#4	369,994	369,907	<1	F#3	+7,5	185,800	185,808	+8	Male
24	+2	N/A	C5	523,251	523,262	<1	C5	+15	527,804	527,422	+14	Female
25	+3	N/A	A4	440,000	440,108	<1	A3	+22,5	222,878	222,842	+22	Male
26	+4	N/A	D#4	311,127	311,208	<1	D#4	+30	316,565	316,969	+32	Female
27	+5	N/A	C4	261,626	261,557	<1	C3	+37,5	133,677	133,645	+37	Male
28	+6	N/A	F#4	369,994	369,907	<1	F#4	+45	379,737	379,905	+46	Female
29	Control	Minor 2nd	A4	440,000	440,061	<1	A#4	+100	466,164	466,275	+100	Female
30	0	N/A	F#4	369,994	369,907	<1	F#3	0	184,997	184,97	<1	Male
31	-1	N/A	A#4	466,164	466,073	<1	A#3	-7,5	232,074	231,944	-8	Male
32	-2	N/A	A4	440,000	440,082	<1	A4	-15	436,204	436,387	-14	Female
33	-3	N/A	C4	261,626	261,557	<1	C3	-22,5	129,124	129,079	-23	Male
34	-4	N/A	D#4	311,127	311,208	<1	D#4	-30	305,782	306,101	-28	Female
35	-5	N/A	F#4	369,994	369,907	<1	F#3	-37,5	181,033	181,056	-37	Male
36	-6	N/A	A4	440,000	440,082	<1	A4	-45	428,710	428,798	-45	Female
37	-6	N/A	F#4	369,994	369,907	<1	F#4	-45	360,501	360,415	-45	Female
38	+5	N/A	A#4	466,164	466,073	<1	A#3	+37,5	238,186	238,09	+37	Male
39	-1	N/A	A4	440,000	440,108	<1	A3	-7,5	219,049	219,126	-7	Male
40	-4	N/A	F#4	369,994	369,907	<1	F#3	-30	181,819	181,767	-30	Male
41	Control	Augmented 4th	C4	261,626	261,557	<1	F#4	+600	369,994	369,975	+600	Female
42	0	N/A	A4	440,000	440,108	<1	A3	0	220,000	219,992	<1	Male
43	+1	N/A	F#4	369,994	369,907	<1	F#4	+7,5	371,600	371,331	+6	Female
44	+2	N/A	A4	440,000	440,108	<1	A3	+15	221,914	221,846	+14	Male
45	+3	N/A	C5	523,251	523,262	<1	C5	+22,5	530,096	530,073	+22	Female
46	+4	N/A	A#4	466,164	466,073	<1	A#3	+30	237,156	237,206	+30	Male
47	+5	N/A	A4	440,000	440,082	<1	A4	+37,5	449,635	449,410	+37	Female
48	+6	N/A	C5	523,251	523,262	<1	C5	+45	537,030	537,676	+47	Female
49	Control	Minor 3rd	F#4	369,994	369,907	<1	A4	+300	440,000	440,083	+300	Female
50	0	N/A	C4	261,626	261,557	<1	C3	0	130,813	130,816	<1	Male
51	-1	N/A	D#4	311,127	311,208	<1	D#4	-7,5	309,782	309,873	-7	Female
52	-2	N/A	A4	440,000	440,108	<1	A3	-15	218,102	218,111	-15	Male
53	-3	N/A	F#4	369,994	369,907	<1	F#4	-22,5	365,216	365,122	-23	Female
54	-4	N/A	C5	523,251	523,262	<1	C5	-30	514,262	514,771	-28	Female
55	-5	N/A	A#4	466,164	466,073	<1	A#3	-37,5	228,088	228,040	-38	Male
56	-6	N/A	C4	261,626	261,557	<1	C3	-45	127,457	127,413	-46	Male
57	+3	N/A	F#4	369,994	369,907	<1	F#4	+22,5	374,834	374,600	+21	Female
58	+2	N/A	D#4	311,127	311,208	<1	D#4	+15	313,834	313,882	+15	Female
59	0	N/A	A#4	466,164	466,073	<1	A#3	0	233,082	233,094	<1	Male
60	-3	N/A	C5	523,251	523,262	<1	C5	-22,5	516,495	517,383	-20	Female
61	-2	N/A	A#4	466,164	466,073	<1	A#3	-15	231,071	231,108	-15	Male
62	-3	Control for 114	A4	440,000	440,108	<1	A3	-22,5	217,159	217,084	-23	Male
63	-5	Control for 116	C4	261,626	261,557	<1	C3	-37,5	128,010	128,014	-37	Male
64	+1	Control for 102	A#4	466,164	466,073	<1	A#3	+7,5	234,094	234,079	+7	Male
65	0	Control for 111	D#4	311,127	311,208	<1	D#4	0	311,127	311,076	<1	Female
66	+2	Control for 133	C5	523,251	523,262	<1	C5	+15	527,804	527,422	+14	Female
67	+3	Control for 134	A4	440,000	440,108	<1	A3	+22,5	222,878	222,842	+22	Male
68	+5	Control for 136	C4	261,626	261,557	<1	C3	+37,5	133,677	133,645	+37	Male
69	+6	Control for 137	F#4	369,994	369,907	<1	F#4	+45	379,737	379,905	+46	Female
70	+4	Control for 165	A#4	466,164	466,073	<1	A#3	+30	237,156	237,206	+30	Male
71	-4	Control for 145	D#4	311,127	311,208	<1	D#4	-30	305,782	306,101	-28	Female
72	-2	Control for 173	A4	440,000	440,108	<1	A3	-15	218,102	218,111	-15	Male
73	-1	Control for 172	D#4	311,127	311,208	<1	D#4	-7,5	309,782	309,873	-7	Female
74	-6	Control for 177	C4	261,626	261,557	<1	C3	-45	127,457	127,413	-46	Male

3.2 Apparatus

An HP laptop was used to run the software. The operating system was Microsoft Windows 10. The participants listened to the sounds through a pair of M-Audio HDH50 headphones. The sound level was set to 50 out of 100 on the Windows' mixer. The input device was a wireless Dell mouse. To prevent any possible disruptions, the wi-fi connectivity, the anti-virus software, and the firewall software were disabled on the computer.

A computer-based surveying tool was written using the Java Development Kit by Assoc. Prof. Dr. Barış Bozkurt, a key member of the aforementioned research group. The user interface of the software was designed in two columns. On the left column, 4 buttons were located one under the other. On the right column, there were two checkboxes one under the other, along with an information box located below them (Figure 3.1).

Figure 3.1 : The software interface.

The first button on the left read "Basla(.jar dosyasini seciniz)" [Start(select the .jar file)] and was used by the author of this text to load the archive file with the extension ".jar" that was necessary by the design of the software to run the survey. The one below read "Dinle" [Listen] and it was used by the participants to start the survey and replay the sounds when needed.

The third button on the left side read "Karari kaydet" [Save the decision] and was employed by participants to finalize their decision made using the checkboxes located on the right column. Thus, this button would take no effect if none of the checkboxes were clicked.

The final button located at the bottom of the left column was “Sonrakine gec” [Next]. The function of this button was to listen the next question after a decision regarding the question at hand was finalized. Thus, the button would work only after the “Sonrakine gec” button was pressed.

On the right-hand side, the top of the column was reserved for the heading “Karar” [Decision], referring to the two checkboxes below, one reading “DOGRU” [Correct] and the other reading “YANLIS” [Incorrect].

At the bottom of the left column located was the information box. This box would guide the participants about what they should do to proceed. At the beginning of the survey, for example, it would inform participants about the fact that the survey would begin once they clicked “Dinle” button. Similarly, when all questions were answered, the box would state that the survey was completed. If the participant would click “Karari kaydet” without checking a box or would click “Sonrakine gec” without clicking “Karari kaydet”, the box would inform them that they’ve missed a step mandatory to proceed in the survey.

After all questions were answered, the software would automatically produce a file with the extension “.txt” with the participants’ answers saved in it.

3.3 Open-Ended Questions

Following the computer-based survey, conversations were made with the participants to ask them some open-ended questions that would give the chance to put the data gathered using the survey software in some context. In short, these questions were designed to control the effects of some additional variables that shape the behavior patterns of the jury members.

Along with the rather standard questions regarding gender and age, some additional questions were asked: Content of their formal education on music, how long they’ve been teaching in the university (if they teach), if they ever served as a jury member at the conservatory entrance exams, and if they had absolute pitch. Additionally, they’ve been asked if they had any other comments.

The data gathered through these open-ended questions were not employed in this thesis in their full extent, since they were beyond the scope of the thesis. Yet it is believed that these data might prove useful in the future studies with a more extended scope.

3.4 Participants

Since the main concern was to get information about the acceptable ranges for jury members of the conservatory entrance exam regarding the single note repetition performances, and since these juries consist of faculty with various musical backgrounds, the only criterion when picking participants was to make sure they are either faculty members teaching music-related topics at the university, or they are or soon will be eligible to become faculty members.

A total number of 35 participants (24 male, 11 female) participated in the data gathering operation. It should be noted that the number of participants from each gender are not well distributed. Any possible effects of gender were not among the primary concerns at that moment, and thus gender was not an important criterion when the participants were picked.

The average age of these 35 participants was roughly 30 with youngest ones being 20 and the oldest one being 61 years old. It should be noted that the age differences among the participants are not well distributed. Any possible effects of age were not among the primary concerns at that moment, and thus age was not an important criterion when the participants were picked. None of the participants reported any health issues regarding their auditory system.

11 of these participants were actively teaching at the university, 12 of them were eligible to teach even though they didn’t teach at the time of the survey (meaning that they had at least a bachelor’s degree in a music-related field), and the remaining 12 of them were pursuing their bachelor’s degree which gives them the authorization to teach at a university.

Most of the participants (33, to be exact) were of Turkish nationality, and two of them were expats. This variable was also overlooked, since both expats comply with the criterion for them to be teaching at the university.

Five of the participants served as a jury member at the conservatory entrance exams at least once. Two of them had absolute pitch. Four of them had a Turkish music-oriented education. Another four of them had a Western art music-oriented education. The remaining 27 of them had an education with the mixture of both Turkish and Western musics.

An informed consent form was read and signed by each of the participants, informing them about why there was a need for the survey at hand, how the surveying process works, and how the data gathered from the survey will be employed. Since the survey was conducted in an environment where both Turkish and English are being spoken, the consent form existed in two different versions, one in Turkish, and one in English (See appendices A.1 and A.2).

Four of the participants were left out when the results of the survey were analyzed, since they failed to evaluate the control questions as unsuccessful, even though those questions contained performances that deviated from the pitches given by piano sounds more than 100 cents. This decision left 31 participants to analyze.

Even though it would be interesting to observe the possible effects of having absolute pitch on the surveyors' behavior patterns, the lack of an enough number of participants identifying themselves as absolute pitch owners made it impossible to observe any such effect. It should be noted, however, that both self-described absolute pitch owners in the original 35 participants were also part of the final 31 surveyors taken into account in the statistical analysis.

3.5 Procedure

The participants were seen on a predetermined day and time. The location the survey was taken was different for each participant. This might seem to be an erroneous way to conduct a survey, since the location might alter the behavior of the participants, but since a soundproof area that's open for use all day long in the campus was not available, this was necessary because of practical reasons. To compensate for this faulty conduct, it has been tried to make sure that each participant was at ease when taking the survey by giving them as much time as they want, letting them ask questions as much as they want, and providing them bottled water to drink.

To make sure that the participants don't lose focus about our question regarding if they'd give a passing grade to the performances they hear through the headphones, two notes, one in English and one in Turkish, (see appendices B.1 and B.2) were put on the keyboard of the laptop, with the hope that this will remind them that this is not a test to evaluate their skill to discriminate pitches, but a survey to understand their personal choices when evaluating the single note repetition performances.

The participants were asked to read the informed consent forms mentioned above. After they read the text, they were asked if they had any questions regarding the survey and survey-related topics (such as the Turkish conservatory entrance exams in general) and these questions got answered carefully and thoroughly. It was also stated that following the survey they'd be asked a small number of open-ended questions to get some demographic data to employ in our study. Only after our participants' all questions answered, they were asked to sign the form if they were ready to give their consent to participate.

After consent forms were signed, it was asked to participants to sit in front of the computer and they were taught how the software interface works employing recordings of 5 performances that was not included in the final question set of the survey. Since the interface was in Turkish, special attention was paid to expats to make sure that they had no confusion about how it worked.

After making sure that the participant understood how the interface worked, the survey started using software with the real question set. After the participants answered all 74 questions, the open-ended questions were asked to them to finalize the session.

3.6 Tools for Data Analysis

The data gathered through the survey was analyzed with the help of the software named IBM SPSS Statistics.

To see if the questions in the survey agree in their judgements regarding the participants, we've used the internal consistency measure called Cronbach's alpha (Mitchell and Jolley, 2013, p. 124). This measure is used to provide evidence regarding the agreement among the "individual items" in a test. If measure gives a coefficient above 0,70, then the internal consistency of the test is "at least respectable", and if it is above 0,90, then it is "excellent" (Mitchell and Jolley, 2013, p. 124).

To depict the correlation between the success rates of deviation regions and the (0) region, we've used the correlation coefficient called the Pearson *r* (Mitchell and Jolley, 2013, p. 256).

Since we needed to see if the success rates differed between the groups (regions) in a significant way, some post hoc tests were needed. To be able to do those tests, first the ANOVA (analysis of variance) test was done to see if the difference among the group means were statistically significant. The *p* value resulted in the ANOVA test had to be lower than 0,05 to decide that such a difference was significant (Mitchell and Jolley, 2013, pp. 400, 446).

If the ANOVA test reported some significance, reporting that there’s indeed some difference among the group means, then the post hoc tests, were conducted (Mitchell and Jolley, 2013, pp. 446, 700). Two different post hoc tests were employed: LSD (least significant difference), and Bonferroni. Since the former is a liberal test (prone to errors), and the latter is a conservative one (lacking statistical power) (Field, 2013, pp. 458-459), a comparison of their results was helpful for us to interpret the results.

4. FINDINGS OF THE SURVEY

4.1 Preliminary Analyses and Results

When the answers given by the 31 participants were categorized with respect to their deviation regions and fed into Cronbach’s Alpha test to check internal consistency of the survey, the coefficient was 0,971, meaning that the internal consistency of the survey was quite high.

When the questions from the (0) region got compared with the questions from the (+) deviation regions regarding their success rates by the help of Pearson correlation test, the result was *r*=-0,425, with the two-tailed significance test coefficient being 0,0, meaning that there’s significant correlation between the compared regions (Table 4.1).

A comparison between the questions from the (0) region and the questions from (-) deviation regions gave a similar outcome, with Pearson correlation test producing the result of *r*=0,243, and the two-tailed significance test coefficient being again 0,0, meaning that there’s significant correlation between the compared regions (Table 4.1).

Table 4.1 : Correlations between success rates and regions.

		Success Rate
(0) and (+) regions	**Pearson Correlation**	-0,425
	Significance (2-tailed)	0,0
	Number of items	1178

		Success Rate
(0) and (-) regions	**Pearson Correlation**	0,243
	Significance (2-tailed)	0,0
	Number of items	1178

Although both are significant, the difference between the results of the aforementioned correlation tests raises the question whether the participants were less tolerant to the unsuccessful performances sung higher than the reference piano pitch compared to the unsuccessful performances sung lower than the reference piano pitch.

A graphical representation of the success rates for reach region can be seen on Figure 4.1. Y-axis represents success rates, while x-axis represents the regions. Green line marks the (0) region, and the red line marks 60% threshold for success rates, which is the minimum score necessary for a prospect student in the exams to pass the first phase.

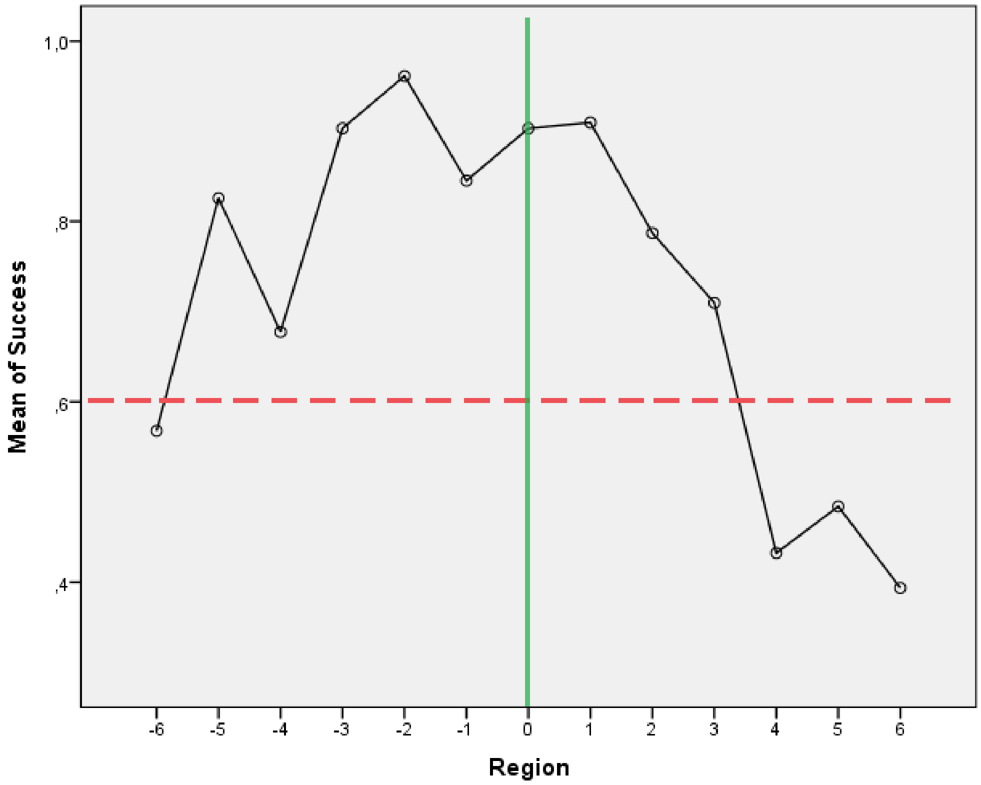

Figure 4.1 : The rates of success for each region.

More details in the shape of descriptive statistics on the rates of being found successful by the participants for reach region can be found on Table 4.2.

Table 4.2 : Descriptive statistics on the rates of being found successful by the participants for reach region.

Descriptives

Success

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum	Between-Component Variance
						Lower Bound	Upper Bound			
-6		155	,57	,497	,040	,49	,65	0	1	
-5		155	,83	,381	,031	,77	,89	0	1	
-4		155	,68	,469	,038	,60	,75	0	1	
-3		155	,90	,297	,024	,86	,95	0	1	
-2		155	,96	,194	,016	,93	,99	0	1	
-1		155	,85	,363	,029	,79	,90	0	1	
0		248	,90	,296	,019	,87	,94	0	1	
1		155	,91	,288	,023	,86	,96	0	1	
2		155	,79	,411	,033	,72	,85	0	1	
3		155	,71	,455	,037	,64	,78	0	1	
4		155	,43	,497	,040	,35	,51	0	1	
5		155	,48	,501	,040	,40	,56	0	1	
6		155	,39	,490	,039	,32	,47	0	1	
Total		2108	,73	,444	,010	,71	,75	0	1	
Model	Fixed Effects			,403	,009	,71	,75			
	Random Effects				,055	,61	,85			,037

As one can see on Figure 4.1, the somewhat steady trend of lowering rates of being found successful as one goes gradually from the region (0) into region (+6) gets interrupted with a small bump at the (+5) region. Moreover, a similar but bigger bump interrupting the expected trend also occurs on the (-5) region.

To find the underlying cause of these irregularities at the (-5) and (+5) regions, we've compared region (0)'s success rates with the other regions' success rates by the means of the ANOVA test, resulting with a significance score of 0,0, meaning that the difference between the group means is highly significant.

The ANOVA was followed by post hoc tests (the LSD, and the Bonferroni) to see the details about the differences between regions. The regions of (+3), (+4), (+5), and (+6) are evaluated as significantly different from the region (0) by both tests. According to LSD test, (+2) region is also significantly different (Table 4.3).

Table 4.3 : Comparison of region (0)'s success rates with the other regions'.

Post Hoc Tests

Multiple Comparisons

Dependent Variable: Success

	(I) Region	(J) Region	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
LSD	0	-6	,335*	,041	,000	,25	,42
		-5	,077	,041	,061	,00	,16
		-4	,226*	,041	,000	,14	,31
		-3	,000	,041	1,000	-,08	,08
		-2	-,058	,041	,159	-,14	,02
		-1	,058	,041	,159	-,02	,14
		1	-,006	,041	,876	-,09	,07
		2	,116*	,041	,005	,04	,20
		3	,194*	,041	,000	,11	,27
		4	,471*	,041	,000	,39	,55
		5	,419*	,041	,000	,34	,50
		6	,510*	,041	,000	,43	,59

*. The mean difference is significant at the 0.05 level.

Dependent Variable: Success

	(I) Region	(J) Region	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Bonferroni	0	-6	,335*	,041	,000	,19	,48
		-5	,077	,041	1,000	-,06	,22
		-4	,226*	,041	,000	,08	,37
		-3	,000	,041	1,000	-,14	,14
		-2	-,058	,041	1,000	-,20	,08
		-1	,058	,041	1,000	-,08	,20
		1	-,006	,041	1,000	-,15	,13
		2	,116	,041	,383	-,02	,26
		3	,194*	,041	,000	,05	,33
		4	,471*	,041	,000	,33	,61
		5	,419*	,041	,000	,28	,56
		6	,510*	,041	,000	,37	,65

*. The mean difference is significant at the 0.05 level.

When it comes to (-) regions, however, the picture is much complex and puzzling. According to both tests, there's indeed a significant difference between the (0) region and the regions of (-4) and (-6). Yet the missing region of (-5) inhibits us to deduce that the range we've been searching is between (Table 4.3).

Thus, it was decided that at this phase of the research, according to preliminary results, the range we seek might be between the regions (-6) and (+4), disregarding the significant differences at the regions (-4), (+2), and (+3), since they have rather low mean difference values (Table 4.3).

That being the case, the following preliminary conclusion took form: When the participants took the role of conservatory jury members to evaluate the performances of single note repetitions as successful, the tolerable range for frequency deviations from the aimed pitch given by the means of a note struck on piano is located between the regions (+4) and (-6). In other words, it is located roughly between +30 cents and -45 cents of deviation from the target pitch.

4.2 Updated Data

Interestingly, as one can see on Figure 4.1, some deviation regions were considered more successful by the participants than (0) region which, of course, contains all the performances that don't deviate from the aimed pitch according to the measurements of the software Tony. The most apparent example of this among all the deviation regions is (-2) region.

This led to the suspicion that there was an error in the survey material an we've realized that one original performance recording (sung by a female) employed in the question set had relatively low success rates in every region, including the (0) region where every performance is successful according to the measurements of Tony software. While other performances from the (0) region had the success rates between 90% and 100%, this one flawed performance recording had a success rate even lower than 75% in the (0) region.

Interestingly, when this problematic recording was analyzed in Tony, it was seen that according to Tony, it contained only a successful performance of a single pitch. Thus, the realization came that there wasn't a human error when picking this recording to be employed in the question set. It fitted the criteria to be containing only a single pitch. Yet when the author of this thesis subjectively interpreted the spectrogram of that recording by comparing it with the spectrograms of other recordings, it was seen that attack and decay sections of the pitch envelope contained relatively wide glissandos, and the part where a steady pitch got sustained was quite short (Figure 4.2).

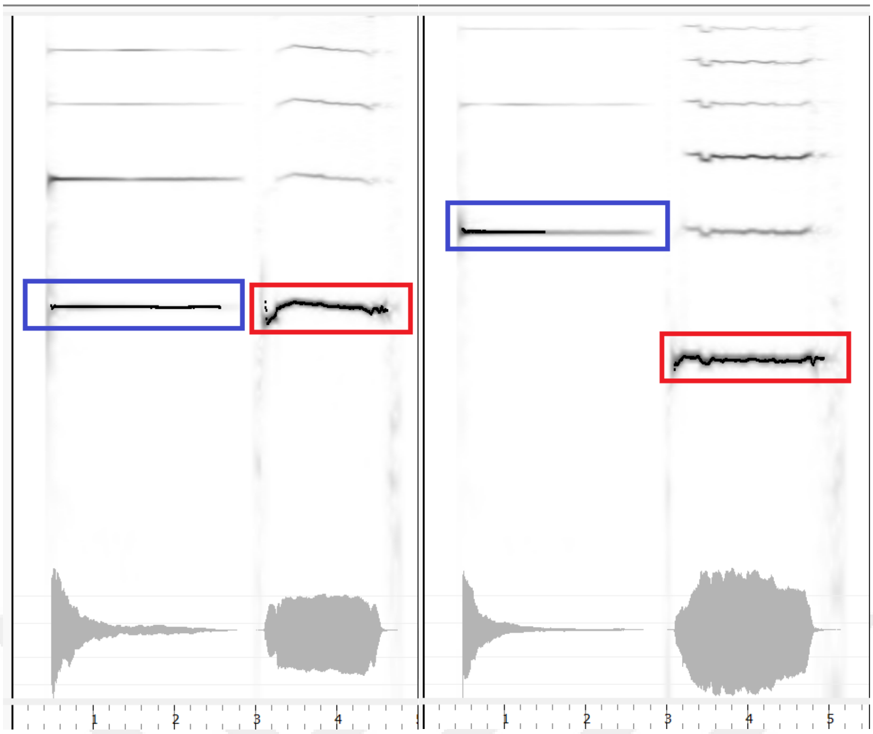

Figure 4.2 : The problematic recording (on the left) compared with another.

Figure 4.2 depicts the problematic recording (on the left) side by side with another recording used in the question set (on the right). Blue boxes mark reference sounds, and red boxes mark the performances.

This situation led to the conclusion that the employment of Tony as a single source of information regarding pitch will not suffice in studies like this thesis. A performance that seems successful according to the measurements done on the computer can yield otherwise when it gets listened by the human ears.

4.3 Analyses and Results Based on the Updated Data

Thus, the rotten egg was from the data, and the tests were redone. Now, when the questions from (0) region got compared with the questions from the (+) deviation regions regarding the rates of being found successful by the participants by the means of Pearson correlation test, the result was *r*=-0,479, with the two-tailed significance coefficient being 0,0. When the questions from (0) region got compared with the questions from the (-) deviation regions regarding the rates of being found successful by the participants by the means of Pearson correlation test, the result was *r*=0,322, with the two-tailed significance coefficient being 0,0. Hence, both correlations were significant (Table 4.4).

Table 4.4 : An update on Table 4.1 after the problematic recording was removed.

		Success Rate
(0) and (+) regions	Pearson Correlation	-0,479
	Significance (2-tailed)	0,0
	Number of items	1054

		Success Rate
(0) and (-) regions	Pearson Correlation	0,322
	Significance (2-tailed)	0,0
	Number of items	961

It can be confidently said that when the problematic recording was removed from the pool, the gap between the correlation coefficients of two deviation directions' rates of being found successful compared to the (0) region got smaller. The details regarding the results of the analysis exercised on the updated data can be found on Figure 4.3 and Table 4.5.

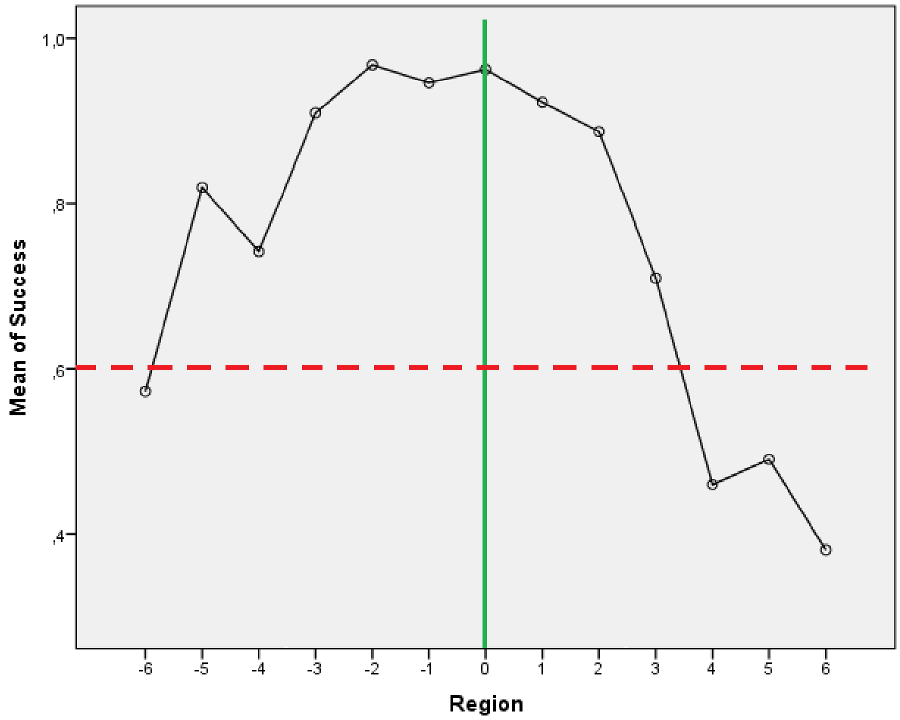

Figure 4.3 : An update to figure 4.1 after the problematic recording was removed.

Table 4.5 : An update to table 4.2 after the problematic recording was removed.

Descriptives

Success

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum	Between-Component Variance
						Lower Bound	Upper Bound			
-6		124	,57	,497	,045	,48	,66	0	1	
-5		155	,82	,386	,031	,76	,88	0	1	
-4		93	,74	,440	,046	,65	,83	0	1	
-3		155	,91	,288	,023	,86	,96	0	1	
-2		155	,97	,177	,014	,94	1,00	0	1	
-1		93	,95	,227	,024	,90	,99	0	1	
0		186	,96	,191	,014	,93	,99	0	1	
1		155	,92	,268	,022	,88	,97	0	1	
2		124	,89	,318	,029	,83	,94	0	1	
3		155	,71	,455	,037	,64	,78	0	1	
4		124	,46	,500	,045	,37	,55	0	1	
5		155	,49	,502	,040	,41	,57	0	1	
6		155	,38	,487	,039	,30	,46	0	1	
Total		1829	,75	,430	,010	,73	,77	0	1	
Model	Fixed Effects			,380	,009	,74	,77			
	Random Effects				,060	,62	,89			,045

On Figure 4.3, y-axis represents success rates, while x-axis represents the regions. The dark green line marks the (0) region, and the red line marks 60% threshold for success rates, which is the minimum score necessary for a prospect student in the exams to pass the first phase. Table 4.5 shows descriptive statistics on the rates of being found successful by the participants for each region.

The ANOVA test was redone employing updated data and resulted the same significance score as before, 0,0, meaning that the difference between the group means is highly significant.

Post hoc analyses of the updated data also yielded similar results with the earlier tests, but with even more defined thresholds. Table 4.6 depicts the comparison of region (0)'s success rate with the other regions' by the means of post hoc analyses; the LSD test result on top, the Bonferroni test result on bottom.

Table 4.6 : An update to table 4.3 after the problematic recording was removed.

Post Hoc Tests

Multiple Comparisons

Dependent Variable: Success

	(I) Region	(J) Region	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
LSD	0	-6	,390*	,044	,000	,30	,48
		-5	,143*	,041	,001	,06	,22
		-4	,220*	,048	,000	,13	,31
		-3	,053	,041	,202	-,03	,13
		-2	-,005	,041	,896	-,09	,08
		-1	,016	,048	,738	-,08	,11
		1	,040	,041	,335	-,04	,12
		2	,075	,044	,087	-,01	,16
		3	,253*	,041	,000	,17	,33
		4	,503*	,044	,000	,42	,59
		5	,472*	,041	,000	,39	,55
		6	,582*	,041	,000	,50	,66

*. The mean difference is significant at the 0.05 level.

Dependent Variable: Success

	(I) Region	(J) Region	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Bonferroni	0	-6	,390*	,044	,000	,24	,54
		-5	,143*	,041	,042	,00	,28
		-4	,220*	,048	,000	,06	,39
		-3	,053	,041	1,000	-,09	,19
		-2	-,005	,041	1,000	-,15	,14
		-1	,016	,048	1,000	-,15	,18
		1	,040	,041	1,000	-,10	,18
		2	,075	,044	1,000	-,08	,23
		3	,253*	,041	,000	,11	,39
		4	,503*	,044	,000	,35	,65
		5	,472*	,041	,000	,33	,61
		6	,582*	,041	,000	,44	,72

*. The mean difference is significant at the 0.05 level.

It should firstly be noted when dealing with the updated data, LSD and Bonferroni tests completely agree on the thresholds. According to both tests, the regions (-6), (-5), (-4), (+3), (+4), (+5), and (+6) are significantly different from the region (0) (Table 4.6).

This confirmed our earlier conclusion that when the participants took the role of conservatory jury members to evaluate the performances of single note repetitions as successful, the tolerable range for frequency deviations from the aimed pitch given by the means of a note struck on piano is located between the regions (+4) and (-6), disregarding the significant difference at the region (+3), since it has a rather low mean difference value. In other words, the range we seek is located roughly between +30 cents and -45 cents of deviation from the target pitch.

4.4 Effects of Performers' Gender on the Success Rates

Even though the range we seek is established based on the updated data, the reason forced us to update the data in the first place introduced us a much more speculative discussion space regarding the factors affecting the evaluators' behavior besides the pitch itself.

While it's possible to check the effects of countless such factors, employing the data we already have seemed convenient as a preliminary device for discussion. Thus, it was decided to see if the performer's gender had any effects on the participants' behavior when evaluating performers.

It should be noted here that besides the apparent timbral qualities of each genders' voice, there is also the aforementioned difference of singing octave: Males reproduce what they hear one octave lower than the original reference pitch, while females reproduce it as it is.

As a first step to see if the performer's gender had any effects on the participants' behavior when evaluating performers, the Pearson correlation test was conducted to see if there is a correlation between the performer's gender and the success rate. A highly significant but small correlation was found between gender and the success rate, favoring the female performers with the coefficient of 0,076 and the two-tailed significance value of 0,001.

Following this suggestive result, ANOVA tests were run on each region to see if there were significant differences between the success rates among the regions for each gender (Table 4.7).

Table 4.7 : Differences between the success rates among the regions for each gender.

	Descriptives						ANOVA	
Region	Total N	Total Success Rate	Female N	Female Success Rate	Male N	Male Success Rate	F for Success Rate Differences between Genders	Significance
All regions	1829	75%	713	80%	1116	73%	10,51	0,001
+6	155	38%	93	52%	62	18%	20,232	0,0
+5	155	49%	31	87%	124	40%	25,936	0,0
+4	124	46%	31	90%	93	31%	43,768	0,0
+3	155	71%	62	92%	93	57%	25,375	0,0
+2	124	89%	93	98%	31	61%	40,714	0,0
+1	155	92%	62	100%	93	87%	9,067	0,003
0	186	96%	62	100%	124	94%	3,67	0,057
-1	93	95%	0	-	93	95%	-	-
-2	155	97%	62	98%	93	96%	0,855	0,357
-3	155	91%	62	85%	93	95%	3,827	0,052
-4	93	74%	62	68%	31	87%	4,136	0,045
-5	155	82%	31	48%	124	90%	35,912	0,0
-6	124	57%	62	34%	62	81%	35,114	0,0

Table 4.7 depicts a summary of the tests run on the data to see if there were significant differences between the success rates among the regions for each gender. When all regions are considered together, the difference seems significant (0,001), suggesting that the participants favored female voices when evaluating performances.

When only the region (0) is considered, however, the significance disappears (0,057), suggesting that when a performance doesn't have any apparent deviation in the pitch domain from the reference pitch, the performer's gender doesn't have any effects on the evaluation process.

Nevertheless, in all of the (+) regions, as well as in the (-4), (-5), and (-6) regions, there's a significant difference between the success rates of two genders. Among these regions, while female performers are the ones with the higher success rates in the (+) regions, male voices seem to be favored by the participants in the (-) ones.

It should that number of female performers evaluated is smaller than the male performers evaluated. While even in the preliminary data this difference was already present, after the deletion of the problematic female performer from the data, this difference became much more pronounced. When dealing with the (-1) region, for example, this difference proved to be quite problematic, since after the deletion of the faulty performer form the data, no female performers were left in the region. Hence, it was impossible to speculate about the effects of gender for this region.

5. CONCLUSION

Two main outcomes arose from the data presented in the previous section.

5.1 Range(s)

Firstly, as expected, there was indeed a frequency range for a pitch sung by a prospect student to be considered a successful repetition of the reference pitch played on the piano. The survey participants taking the role of the jury members at conservatory entrance exams accepted performances with a quite big range of frequency deviations from the reference pitch as successful.

This range was approximately 75 cents – three fourths of a semitone, that is. A somewhat unexpected result was that this range of 75 cents was not distributed evenly around the absolute correct pitch. For the deviations lower than the absolute correct pitch, the rates of success did not fall under the 60% threshold (the minimum score necessary for a prospect student in the exams to pass the first phase) until the deviations got approximately as big as 45 cents. For the deviations higher than the absolute correct pitch, the rates of success did not fall under the 60% threshold until the deviations got approximately as big as 30 cents.

This asymmetrical distribution of the acceptable range around the absolute correct pitch is not mentioned in the existing literature, yet it seems quite interesting. Surely, it is needed to conduct more surveys and experiments to make sure that such an uneven distribution really exists, but at this point, nevertheless, it is possible to speculate about why such a phenomenon occurs.

One idea that comes to mind is the fact that while the equal temperament is considered to be the standard way of tuning the instruments nowadays, people still are exposed to other temperaments on a daily basis.

In Turkish Makam music, for example, a largely accepted opinion is the idea that a note such as Segah doesn’t refer to a set frequency or doesn’t constitute a set interval with other notes, but refers to a band of different acceptable pitches, depending on context. It can be argued then that for those who has experience with Makam music as a listener or a performer, maybe an F0 difference between two notes might not always constitute two different pitches, but simply a repetition of the same pitch.

It is known that in the case of the well-known Western major scale, most pitches of the equally tempered major scale are quite higher than its just intonated counterpart. If the commonly accepted opinion regarding people's tendency to use just intonation when it's possible is to be taken as a fact, then it is possible to argue the following: Since people deal with both equally tempered and just intonated major scales on daily basis, they might tend to read a band of different pitches (higher pitches of equal temperament compared to just intonation, and lower pitches of just intonation compared to equal temperament) all referring to a single note.

If this theory was taken to be correct, then the result that some of the repetitions of the reference pitches being lower than the reference pitches themselves were considered to be successful repetitions by the participants would make sense, since such repetitions would simply be the ones that tended to just intonation domain in the context of the equal temperament, which was employed in this thesis.

It can be also speculated that the peculiarities of each gender (such as timbre and octave differences), and each performer might have also shaped the range we've found. As can be deducted from the Table 4.7, the difference between the means of each gender gets more significant as the absolute value of the subtraction of the means of each gender gets bigger. Since the means introduced throughout this thesis are weighted averages, this apparent difference between the means of each gender, combined with the considerably fewer number of performances by females compared to males in the updated data, might have distorted our findings regarding the acceptability range and its asymmetrical nature. The uneven distribution of each unique performer among the regions might have further affected this result.

For example, the case of male performers in the (+6) and (-6) regions suggests that the peculiarities of each performance, besides the fundamental frequency, might have affected how the performance evaluated and thus eventually the results regarding the ranges. Even though the fact that both these regions contain deviations of 45 cents, the male performances in the (-6) region got a quite high success rate of 81% (Table 4.7).

It should be clear that this success rate, which is well above the threshold of 60%, is quite distinct compared to the other performances in the (+6) and (-6) regions.

Interestingly, there are two questions performed by male voice in the (-6) region, but they are identical, since one of the questions is simply the control for the other. Thus, it can be argued that effectively there's only one male performance in this region. In the (+6) region, on the other hand, there are two questions performed by male voice, different from each other, but one of them originated from the same recording as the twins in the (-6) region. Curiously enough, this recurring performance which led to a quite high success rate in the (-6) region is the one that got lower score among the two male performances in the (+6) region. In other words, this specific performance was considered quite successful when it was approximately 45 cents lower than the aimed pitch, but it was not considered successful when it was approximately 45 cents higher than the aimed pitch.

For the (+6) region, then, the result regarding this specific performance is in accordance with what was seen so far, but the success rate of this performance in the (-6) region violates the acceptability range we've observed. It is as if the version of the performance in the (-6) region had a specific trait that caused the participants disregard the pitch content and deem it successful, but when the performance was processed to be higher, as in the region (+6), that very same trait worked against it.

Since the question set was designed only with the pitch content in mind, the available data regarding the survey doesn't offer much to make informed guesses on why such a thing happens. Yet, it suggests, nevertheless, that further studies with a much more controlled question set design should follow to see if the speculations produced here have merit.

Other possible effects of the peculiarities (such as timbre and octave differences) of each gender and each performer to the evaluation process are discussed below.

5.2 Other Variables Affecting Evaluators' Judgement

As reported by Kollmeier et al. (2008), JND for frequency is 3 Hz for sine waves below 500 Hz (p. 65). As Table 2.3 suggests, F0 values of nearly all sounds employed in the survey are below this 500 Hz threshold, be them reference pitches on the piano or reproduction of those pitches by voice. Yet as Kollmeier et al. (2008) point out, for complex sounds such as of piano and of voice, the JND for frequency is even smaller: 1 Hz in the range of 80 Hz to 500 Hz (p. 65).

Zarate et al. (2013) also reports that recent studies suggest that both musicians and non-musicians discriminate intervals better when complex sounds are used instead of pure tones (p. 1). They've found that the pitch discrimination threshold was 14,6 cents for musicians and 44,1 cents for non-musicians and except for the interval of 25 cents, the interval-discrimination was best with pure tones (Zarate et al., 2013, p. 3, 5).

The acceptable range of 75 cents for pitch deviations in the performances heard during the survey then is well beyond JND value for human perception. This means that there must be some other variables affecting the perceived pitch. This leads us to the second main outcome of the data: We've found out that attack, decay, sustain, and release (ADSR) of the pitch envelope, timbre, octave differences, and other such variables affected the decision of the participants regarding the quality of the performance in a way we've failed to foresee.

At the start of the designing process of the question set, we've supposed that the voice recordings that contained only a single pitch content that agrees with the pitches of the equal temperament according to the software Tony would be a good start to devise voice recordings tampered to our needs. Yet, it seems that when deciding on the content of a recording, Tony focuses on the sustain section of the pitch envelope where the frequency is somewhat stable, disregarding the other parts of the envelope. Humans, however, it seems, do not disregard the unstable parts of pitch envelope, when evaluating pitch accuracy of a voice performance.

Since ADSR values of pitch envelope will be different for each performance, we suggest that the question set for future studies should be devised from a recording of a single performance with a long and stable sustained single pitch content. This would prevent the entrance of unwelcome pitch envelope variables into the data. If it will be somehow necessary to use more than one performance recording, however, we strongly suggest a much more rigid audition process than we've employed here.

Another option to devise material for future studies would be the use of sound synthesis methods to artificially construct human voice performances, where all the parameters tactfully manipulated to make sure they won't introduce any unwanted variables. Obviously, it might prove quite hard to synthesize such sounds that would replace the real performance recordings, but in some cases the lack of realism might be a good trade off in an attempt to get undistorted and reliable data.

These suggestions, of course, imply a need for studies dealing with a much more different research question than the one investigated here. This question would be regarding the effects of ADSR values of pitch envelope on the jury members' decision at the conservatory entrance exams.

Since it was realized that unique qualities of each performance, such as pitch envelopes, affect how it was evaluated by the jury members, the question of what other such qualities might have been affecting the evaluation processes emerged. Timbre is one of the variables that can be used to depict the complexity of the issue and its potential to be a fruitful research area.

Vurma (2014) reports that recent studies show that timbral differences of the sounds with the same F0 value can be perceived as pitch-shifts by both musicians and non-musicians, and her own research also suggests similar results (pp. 1, 11). Oxemham (2012) also reports that "intensity" (loudness) and "brightness" (timbre) can affect the pitch perceived (p. 13337). Similarly, Zarate et al. (2013) report that recent studies found that the manipulation of timbre (between dull and bright) affects the perception of the melodic interval size (p. 1). They report that "pitch changes seem to be best perceived (regardless of timbre) when F0 changes by at least 4%; the perception of smaller F0 changes is more influenced by timbre changes" (Zarate et al., 2013, pp. 7-8).

These reports suggest that timbral differences might make discrimination of different F0s harder, or might make sounds with the same F0 perceived as different pitches. Since both on piano and especially on voice, the timbral choices are manifold, it can be said that it is safe to speculate that when evaluating prospect students' performances, the jury members of conservatory entrance exams are dealing not only with pitch, but also with other domains affecting the pitch.

This means that at conservatory entrance exams, the jury members' evaluation of the prospect students might not be fair, since it can be argued that timbre and pitch envelope characteristics of voice can be altered through education, and should not constitute a criterion to decide which of the prospect students has an "ear" for music.

As discussed in the findings, another variable affecting the evaluation process seems to the gender of the performer. Here, we’re not talking about the possible discriminative tendencies among the evaluators towards one gender, but simply the facts that each gender has a characteristic timbre, and that males reproduce the pitch referenced one octave lower. The possible effects of timbre already discussed above apply readily in this issue, but the effects of octave difference also deserve attention.

As reported, in all of the (+) regions, as well as in the (-4), (-5), and (-6) regions, the difference between the success rates of two genders are statistically significant, while no effect of performers’ gender was observed in the (0) region. Among the regions where there is a significant difference between the success rates of each gender, female performers scored better in the (+) regions, and male performers scored better in the (-) regions. We want to speculate that this might have caused by the octave differences between male and female performers when reproducing the reference pitch.

For female performers, singing in the (+) regions simply means that they are singing higher than the reference pitch, and vice versa in the (-) regions. For males, however, since they always aim one octave lower than the reference pitch, what they sing will be always lower than the reference pitch regardless of the region corresponding with the frequency deviations in their singing, but depending on the frequency deviations in the performance, the distance will change between the reference pitch and the pitch performed. In other words, while a performance by a female voice can simply be either higher (+) or lower (-) than the reference pitch, a performance by male will be either close (+) or far (-) from the reference pitch. (Figure 5.1).

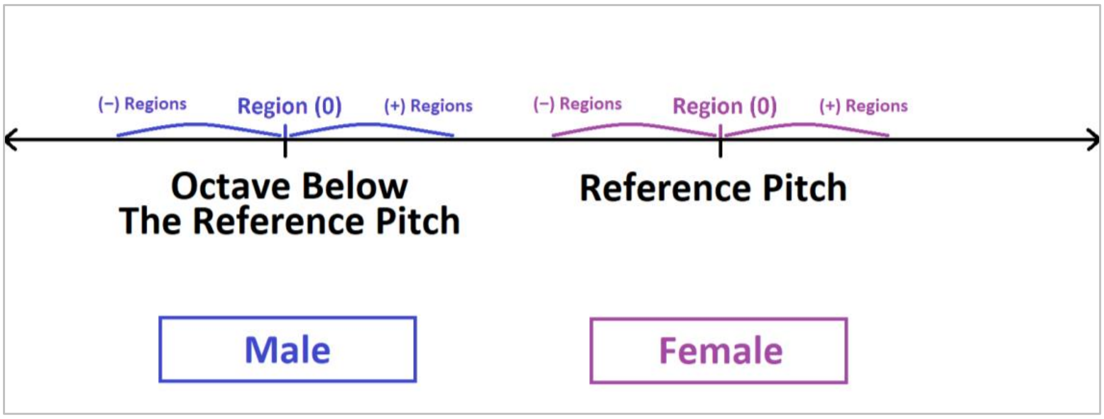

Figure 5.1 : Male performer deviation regions compared with the female performer deviation regions.

As Figure 5.1 suggests, performances by male performers will be always quite remote from the reference pitch regardless of their level of deviation. As mentioned above, in the (-) regions, male performers were favored, and in the (+) regions, female performers were favored. It can be speculated that this also means that in the (-) regions the octave difference of male performers had a positive effect on being counted as successful, and in (+) regions it had a negative effect.

The possible causations for this effect is open to speculation, but since the particularities of each performer such as pitch envelopes, timbre, and gender were overlooked during the design of the survey, focusing only on the pitch content of the performance recordings, it is impossible to conduct further analyses on the available data to make informed guesses. Thus, it is suggested for future studies to choose and/or produce their performance recordings with the consideration of other variables besides the pitch content to design a much more controlled question set that would lead to a much more controlled data set.

Nevertheless, it can be speculated that since the fundamental frequencies of performances by males are one octave lower than the reference pitch, the beatings among the partials of the reference and performed pitches in the case of deviated performances would differ from the beatings occur in the deviated performances by females, since in their case the fundamental frequencies of performance and reference are on the same octave. Thus, the listening experiences of two different performances by each gender might prove quite different from each other, even if the jury members are conscious of these differences.

In short, these findings and speculations suggest that the jury members of the conservatory entrance exams must apply a different kind of listening strategy for each gender, since it is expected from them to disregard the effects of the apparent octave difference and focus on the pitch class relations instead, while evaluating male performers. Then the question comes if it’s viable to trust the jury members to perfect their strategies of listening and evaluating for each gender, or if it’s better to devise some other way of evaluation that removes the variable of performer’s gender would be better.

5.3 Additional Remarks

Wallentin et al. (2010) argue that “the imitation test” such as the one used in Denmark is not suitable for “experimental purposes”, since “since it involves an element of subjective judgment of participants' imitation” (p. 189). Although they do not say that this test, in which participants are asked to “reproduce short rhythmical and melodic phrases using hand-claps and singing” (Wallentin et al., 2010, p. 189), should not be used for other purposes, we believe that the employment of such tests and exams, including the ones in Turkey, should be reconsidered, since it seems like in such tests and exams, many variables, such as timbre, are overlooked, since it’s impossible to control their effects.

The literature regarding binaural hearing was overlooked in this thesis. Since the questions designed for the survey were mono and were introduced to each ear at the same amplitude, and our participants reported no health issues regarding their auditory system, we believe this decision didn’t result in a distorted data. Nevertheless, it is suggested for future studies to take the literature regarding binaural hearing into account to produce more rigorous work. In fact, maybe binaurality might be affecting prospect students of the conservatory entrance exams, since the data entering each of their ear canals differ from the other due to room acoustics.

It should be noted that among many options, we’ve picked equal temperament simply because we needed a common ground and equal temperament being widely used nowadays seemed to offer such a commonality. Still, it should be clear for our readers that we’ve overlooked any possible effects of picking this tuning instead of others. Obviously, tuning is also an important field of discussion in studies like ours. Thus, we acknowledge the fact that our further studies should also deal with the variables added by the issues of tuning.

We’ve used free tools to make sure that any researcher wanting to reproduce the setup of our study to build up on the work done here will be able to do that. Yet, the shortcomings of these free tools rendered, we believe, the process for the preparation of the question set quite inconvenient to the point that probably no one will ever try to reproduce that process ever again. Moreover, the unprecise nature of these free tools caused our material to be much more diverse than we had initially planned, forcing us to invent the concept of deviation regions. As we’ve experienced here, limiting the tools of research solely to unpaid options might seem like a moral highpoint, but sadly such a limit might also limit the potential for academic rigor.

It should be noted that an exhaustive comparison of our survey and Turkish conservatory entrance exams with other surveys, tests, and exams dealing with musical aptitude, skill, talent, expertise, discrimination, apprehension, and such conducted around the world will soon be necessary to get the conservatory entrance exams in Turkey further improved. As our discussion implies, these exams seem to be designed by the instincts of those who have a career in music, but are ignorant of the recent trends and findings in the field. Moreover, they are not well-discussed to be considered as doing what they aim to do. Thus, we argue that these exams inherently possess a risk to be unfair and inefficient to decide which prospect students should be studying at the conservatories.

5.4 Concluding Remarks

Since no similar studies were done on the unique system of the conservatory entrance exams, some decision regarding the design of the survey was left to the instincts. This led to some unforeseen problems and made the progress of the study quite slow. Hereby, some content, such as a section dealing with answers given to the open-ended question, had to be left out of the thesis. Nevertheless, as a first of its kind, the author believes that this thesis is heralding many such studies to come, without reproducing its mistakes. Thus, along with the results of this thesis, it is believed that the mistakes done during the production of this thesis will also guide future studies on the subject.

REFERENCES

Atak Yayla, A. and Yayla, F. (2009). Müziksel algılama ölçeği. *Bildiriler – 8. Ulusal müzik eğitimi sempozyumu*, September 23-25, 2009, Samsun, Turkey. The effect of instrumental timbre on interval discrimination. Retrieved on September 29, 2016, from: http://www.muzikegitimcileri.net/bilimsel/bildiri/samsun/A_Yayla-F_Yayla.pdf

Atar, H. Y., Sayın, A., and Atar, B. (2013). Examination of the Classification Accuracy of Music Education Special Aptitude Exams. *Kuram ve Uygulamada Eğitim Bilimleri*, *(13)*3, pp. 1573-1577. Retrieved on September 29, 2016, from: http://www.kuyeb.com/pdf/en/4bd133db28b1a92a239bcea2b905a87eareng.pdf

Ece, A. S. and Kaplan, S. (2008). Müzik Özel Yetenek Seçme Sınavı'nın Puanlayıcılar Arası Güvenilirlik Çalışması. *Millî Eğitim, 36*(177), pp. 36-48. Retrieved on September 29, 2016, from: http://dhgm.meb.gov.tr/yayimlar/dergiler/Milli_Egitim_Dergisi/177.pdf

Kollmeier, B., Brand, T., and Meyer, B. (2008). Perception of speech and sound. In J. Benesty, M. M. Sondhi, and Y. A. Huang (eds.), *Springer handbook of speech processing* (pp. 61-82). Retrieved on October 26, 2017, DOI: https://doi.org/10.1007/978-3-540-49127-9

Mauch, M., Cannam, C., Bittner, R., Fazekas, G., Salamon, J., Dai, J., Bello, J., and Dixon, S. (2015). Computer-aided melody note transcription using the Tony software: Accuracy and efficiency. *Proceedings of the First International Conference on Technologies for Music Notation and Representation*, May 28-30, 2015, Paris, France. Retrieved on October 28, 2017, from https://code.soundsoftware.ac.uk/attachments/download/1423/tony-paper_preprint.pdf

Mitchell, M. L. and Jolley, J. M. (2013). *Research design explained* (8th ed.). ?: Wadsworth, Cengage Learning.

Oxemham, A. J. (2012). Pitch perception. *Journal of neuroscience, 32*(39), pp. 13335-13338. Retrieved on September 7, 2017, DOI: https://doi.org/10.1523/JNEUROSCI.3815-12.2012

Ullén, F., Mosin, M. A., Holm, L., Eriksson, H., and Madison, G. (2014). Psychometric properties and heritability of a new online test for musicality, the Swedish Musical Discrimination Test. *Personality and individual differences, 63*, pp. 87-93. Retrieved on July 3, 2017, DOI: https://doi.org/10.1016/j.paid.2014.01.057

Vurma, A. (2014). Timbre-induced pitch shift from the perspective of Signal Detection Theory: The impact of musical expertise, silence interval, and pitch region. *Frontiers in psychology, 5*(44). Received on September 7, 2017, DOI: http://dx.doi.org/10.3389/fpsyg.2014.00044

Wallentin M., Nielsen, A. H., Friis-Olivarius, M., Vuust, C., and Vuust, P. (2010). The Musical Ear Test, a new reliable test for measuring musical competence. *Learning and individual differences, 20*, pp. 188-196. Retrieved on July 3, 2017, DOI: https://doi.org/10.1016/j.lindif.2010.02.004

Zarate, J. M., Ritson, C. R., and Poeppel, D. (2013). The effect of instrumental timbre on interval discrimination. *PLoS ONE, 8*(9), pp. 1-9. Retrieved on September 7, 2017, DOI: https://doi.org/10.1371/journal.pone.0075410

Url-1 <*http://www.sengpielaudio.com/Calculations03.htm*>, date retrieved October 26, 2017.

APPENDICES

APPENDIX A.1 : Consent form in English.
APPENDIX A.2 : Consent form in Turkish.
APPENDIX B.1 : Survey annotation in English.
APPENDIX B.2 : Survey annotation in Turkish.

APPENDIX A.1

Master’s Thesis Data Gathering Operation
Informed Consent Form

Institution: Istanbul Technical University Centre for Advanced Studies on Music (ITU MIAM)
Program: Master’s Programme with Thesis on Music (English) (1st Education)
Thesis title: The Acceptable Range(s) for Single Note Repetitions
Thesis advisor: Assoc. Prof. Ozan Baysal, PhD
Researcher’s name: Res. Assist. Oğul Köker
Researcher’s address: İstanbul Teknik Üniversitesi Maçka Kampüsü, Yabancı Diller Yüksekokulu Binası Kat 3, Müzik İleri Araştırmalar Merkezi Oda 124, Şişli, Istanbul, 34367, TURKEY
Researcher’s e-mail address: ogulkoker@yahoo.com
Researcher’s cellphone number: +90 530 305 84 40

Dear participant,

Res. Assist. Oğul Köker, serving at ITU MIAM and working as a scholar in the TÜBİTAK [Türkiye Bilimsel ve Teknik Araştırmalar Kurumu / Turkish Scientific and Technical Research Institute] project titled “Müzik Algısı Ölçme-Değerlendirme Sınavlarında Ses İşleme Araçlarının Kullanılabilirliği” [Applicability of Sound Processing Tools in Music Apprehension Assessment and Evaluation Examinations] (project no. 215K017), is working on the thesis titled “The Acceptable Range(s) for Single Note Repetitions” under the advisorship of Assoc. Prof. Ozan Baysal, PhD.

The goal of the above-mentioned project is the development of assistive software to be used in the conservatory entrance elimination examinations. In the development, it became necessary to gather data for the single pitch repetition analysis tool in order to teach it how to differentiate a good performance from a bad one. With Oğul Köker’s thesis, it’s our aim to gather this data.

To do that, we’ve developed a survey software. The software will make you listen the recordings of some piano reference sounds followed by their reproductions by human voice, and it will ask you to score those performances as passing or failing, as if you were part of the conservatory entrance examinations jury. The data from this survey will be analyzed, and employed for the development of the skill to differentiate a good performance from a bad one in the software that’s being developed as part of the project.

We invite you to support our study by participating our master’s thesis data gathering operation.

This study is done for scientific purposes only and the anonymity of participant information is crucial. Since our aims are to improve the qualification examinations, and to test the employability potential of sound processing tools in the examinations, your identity as a participant doesn’t have any importance for us. Your name won’t be featured in the date we gathered from you, and this data will be kept in a safe environment. If our study succeeds, it’ll be pioneering in the field of creating technologies to support and assist the conservatory examination system, and contributing to a more efficient conduct of the examinations.

The decision to participate in this study is totally optional. If you participate, you have the right to withdraw your consent without giving any explanations in any phase of the study. If you ever want to get any additional information on the study, please contact Oğul Köker of whom the contact information is offered above. Before signing this form, if you have any questions about the study, please ask without any reservation.

If you agree to participate in this study, please sign this form.

I, ... (participant’s name-surname), have read the text above and understood the scope and the aim of the study I’ve been asked to participate, along with the responsibilities of which I’ll be accountable voluntarily. I’ve had the chance to ask questions about the study. I’ve understood that I can stop the process whenever I want, without any obligation to give any explanations, and that I won’t experience any scolding/negativity if I stop. Under these conditions, I agree to participate in the afore-mentioned study with my own will and without any repression and coercion. I took a copy of the form.

Date (DD/MM/YYYY): .... /...../.............

Participant’s name-surname: ....
Participant’s signature: ....
Participant’s cellphone number: ....
Participant’s e-mail address: …………………………………………………

Name-surname of the researcher who gave directions: ....
Signature of the researcher who gave directions: ....

APPENDIX A.2

Yüksek Lisans Tezi Veri Toplama Çalışması
Bilgilendirilmiş Olur Formu

Kurum: İstanbul Teknik Üniversitesi Müzik İleri Araştırmalar Merkezi (İTÜ MIAM)
Program: Müzik Tezli Yüksek Lisans Programı (İngilizce) (1. Öğretim)
Tezin başlığı: The Acceptable Range(s) for Single Note Repetitions [Tek Ses Tekrarları İçin Kabul Edilebilir Aralık(lar)]
Tez danışmanı: Doç. Dr. Ozan Baysal
Araştırmacının adı: Arş. Gör. Oğul Köker
Araştırmacının adresi: İstanbul Teknik Üniversitesi Maçka Kampüsü, Yabancı Diller Yüksekokulu Binası Kat 3, Müzik İleri Araştırmalar Merkezi Oda 124, Şişli, İstanbul, 34367
Araştırmacının e-posta adresi: ogulkoker@yahoo.com
Araştırmacının cep telefonu numarası: 0 530 305 84 40

Sayın Katılımcı,

İTÜ MIAM'da görev yapan ve 215K017 proje numaralı **"**Müzik Algısı Ölçme-Değerlendirme Sınavlarında Ses İşleme Araçlarının Kullanılabilirliği" başlıklı TÜBİTAK projesinde bursiyer olarak çalışan Arş. Gör. Oğul Köker, yukarıda başlığı sunulan yüksek lisans tezini öğretim üyesi Doç. Dr. Ozan Baysal danışmanlığında hazırlamaktadır.

Bahsi geçen proje, konservatuara giriş eleme sınavlarında kullanılmak üzere yardımcı yazılımlar geliştirilmesini amaçlamaktadır. Bu amaç doğrultusunda geliştirilen tek ses icrası analiz yazılımının doğru icrayı yanlış icradan ayırmayı öğrenmesi için veri toplanması ihtiyacı doğmuştur. Oğul Köker'in teziyle, bu verinin elde edilmesi hedeflenmektedir.

Bu hedef doğrultusunda bir anket yazılımı geliştirilmiştir. Yazılım size referans piyano sesleri ve bunların insan sesi tarafından tekrarlarından oluşan ses kayıtlarını dinletecek ve konservatuvar giriş sınavlarında jüri olduğunuz takdirde söz konusu icraya geçer not verip vermeyeceğinizi işaretlemenizi isteyecektir. Bu anketten toplanan veri, tez çalışması kapsamında incelenecek ve proje kapsamında geliştirilen yazılıma doğru icrayı yanlış icradan ayırma yetisi kazandırmakta kullanılacaktır. Araştırmamıza katkı koymanız için sizi yüksek lisans tezi veri toplama çalışmamıza davet ediyoruz.

Bu çalışma bilimsel bir amaçla yapılmaktadır ve katılımcı bilgilerinin gizliliği esas tutulmaktadır. Amacımız eleme sınavlarının kalitesini yükseltilmesi ve sınavlarda ses işleme araçlarının kullanılabilme potansiyelinin sınanması olduğundan katılımcı olarak kimlikleriniz çalışmamız için önem teşkil etmemektedir. Sizden toplanan veride isminiz yer almayacak ve bu veri güvenli bir ortamda saklanacaktır. Çalışmamız başarılı olduğu takdirde, sınav sistemini destekleyici ve kolaylaştırıcı teknolojilerin oluşturulmasının önünün açılması ve sınavların daha etkin gerçekleştirilmesi için bir katkı sunmuş olacaktır.

Bu çalışmaya katılmak tamamen isteğe bağlıdır. Katıldığınız takdirde çalışmanın herhangi bir aşamasında herhangi bir sebep göstermeden onayınızı çekmek hakkına sahipsiniz. Çalışma hakkında ek bilgi almak istediğiniz takdirde lütfen iletişim bilgileri yukarıda sunulmuş olan Oğul Köker ile irtibata geçiniz. Bu formu imzalamadan önce, çalışmayla ilgili sorularınız mevcutsa lütfen çekinmeden sorunuz.

Eğer bu çalışmaya katılmayı kabul ediyorsanız, lütfen bu formu imzalayın.

Ben, ... (katılımcının adı-soyadı), yukarıdaki metni okudum ve katılmam istenen çalışmanın kapsamını ve amacını, gönüllü olarak üzerime düşen sorumlulukları tamamen anladım. Çalışma hakkında soru sorma imkânı buldum. Bu çalışmayı istediğim zaman ve herhangi bir neden belirtmek zorunda kalmadan bırakabileceğimi ve bıraktığım takdirde herhangi bir ters tutum/olumsuzluk ile karşılaşmayacağımı anladım. Bu koşullarda söz konusu çalışmaya kendi isteğimle, hiçbir baskı ve zorlama olmaksızın katılmayı kabul ediyorum. Formun bir örneğini aldım.

Tarih (gün/ay/yıl): ...../...../.............

Katılımcının adı-soyadı: ....

Katılımcının imzası: ....

Katılımcının cep telefonu numarası:..

Katılımcının e-posta adresi:…………………………………………………

Açıklamaları yapan araştırmacının adı-soyadı:...

Açıklamaları yapan araştırmacının imzası:..

APPENDIX B.1

In this survey, we want to learn if you'd give a passing grade to the performances you're hearing, if you were part of the conservatory entrance exam jury.

APPENDIX B.2

Bu ankette, konservatuvar giriş sınavlarında jüri olduğunuz takdirde duyduğunuz icralara geçer not verip vermeyeceğinizi öğrenmek istiyoruz.

CURRICULUM VITAE

Name-Surname: Oğul Köker

Place and Date of Birth: Bornova, İzmir, Turkey; January 14, 1990

E-Mail: ogulkoker@yahoo.com

EDUCATION:

- **B.A.:** 2014, Hacettepe University, Ankara State Conservatory, Department of Musicology

PUBLICATIONS AND PRESENTATIONS:

- Köker, O. and Özçifci, S. (2013). Müzikte Anlam Üzerine Wittgensteincı bir Model [A Wittgensteinian Model on Meaning in Music]. *İstanbul Teknik Üniversitesi Dr. Erol Üçer Müzik İleri Araştırmalar Merkezi Müzik Biliminde Gençler Buluşması 3* [Istanbul Technical University Dr. Erol Üçer Centre for Advanced Studies in Music, Student Conference 3], October 4, 2013, Istanbul, Turkey.
- Köker, O. (2016). Sihirli Flüt ve Masonluk: Bir Yeniden Değerlendirme [Magic Flute and Freemasonry: A Reassessment]. Proceedings from *Hacettepe Üniversitesi Ankara Devlet Konservatuvarı I. Ulusal Müzik ve Sahne Sanatları Sempozyumu* [Hacettepe University Ankara State Conservatory First National Symposium on Music and Performance Arts], May 2-3, 2016, Ankara, Turkey.
- Pala, D., Köker, O., and Charkiolakis., A. (In press). "Shostakovich in Turkey." [To be translated in Russian]. *Dmitry Shostakovich: Researches and materials, Vol. 5*. Moscow: DSCH.