TOWARDS A HYBRID ASSESSMENT MODEL FOR MUSIC CONSERVATORY ENTRANCE EXAMS

OZAN BAYSAL, BARIŞ BOZKURT, TURAN SAĞER, NILGÜN DOĞRUSÖZ

Ozan Baysal1, Barış Bozkurt2, Turan Sağer3, Nilgün Doğrusöz1

1 Istanbul Technical University, Turkey;

2 Universitat Pompue Fabra, Barcelona, Spain;

3 Yıldız Technical University, Turkey.

Abstract

This paper discusses the necessity for employing Music Information Retrieval (MIR) Technologies in Music Conservatory Entrance Examinations. In Turkey, acceptance to a music conservatory is determined through a musical aptitude examination that is usually conducted by a jury committee. While the contents of this exam has become a standard –including mostly questions on pitch recognition and melody/rhythm repetition -, factors such as the amount of time and energy devoted to the exam, differences of assessment criteria between jury members and the usage of limited set of manually constructed question packages (to avoid any leaking of the exam outside) present some shortcomings for a standardized evaluation of applicants. Although there has been a good deal of research made regarding this issue, these researches investigate solely the reliability scores of jury committees, while not making a sound analysis of the applicant performance recordings and comparing them with the jury scores. Our talk will present the findings of such a research project that compares jury scores with performance recordings. At the end we would be proposing a hybrid assessment model, MAST (Musical Aptitude Standard Test), which we believe would significantly contribute to the quality of measurement and evaluation while consuming less resources in music conservatory entrance exams.

Keywords: Conservatory Entrance Exams, Musical Aptitude Tests, Musical Competence, MAST, Musical Aptitude Standard Test, Music Performance Assessment

Introduction

A musical aptitude examination is a general requirement when applying to a music conservatory school. Aiming to test and measure the musical proficiency of an applicant, there are various kinds of different approaches in how to measure musical competence. This paper would present the potential benefits of employing Music Information Retrieval (MIR) Technologies in Music Conservatory Entrance Examinations. In the first part, a brief overview of two main approaches in measuring musical proficiency will be presented; (i) standardized test format, and (ii) jury committee evaluations. Both of the approaches have their own advantages and shortcomings, but in Turkey it is usually the jury committee evaluations that are preferred in music conservatory school exams. The contents of this kind of an examination has become a standard –including mostly questions on pitch recognition and melody/rhythm repetition -, yet since it involves a jury committee, it may present some shortcomings for a standardized evaluation of applicants. Although there has been a good deal of research made regarding this issue, these researches investigate solely the reliability scores of jury committees, while not analyzing the applicant's performance recordings through music/sound technologies and comparing them with jury scores. The second part of the paper will present the findings of such a research project that compares the jury scores with the analysis taken from performance recordings via sound engineering tools. This part will reveal the existence of different assessment criteria between jury members. In addition, it will demonstrate the problem of using limited set of question packages for different applicants (to avoid any leaking of the exam outside). Thus, the scope of this essay is limited with the need of using new technological tools as an aid for the jury committees. At the end we would be proposing a hybrid assessment model, MAST (Musical Aptitude Standard Test), which we believe would significantly contribute to the quality of measurement and evaluation while consuming less resources of time and energy in the musical hearing portion of the conservatory entrance. Our goal is to present supporting and practical mechanisms in order to make the exams as efficient as possible.

Musical Aptitude Tests & Music Conservatory School Examinations

One can categorize the methods of measuring musical proficiency during a music conservatory school examination under two main headings;

Standardized tests that are used to determine various dimensions of aural ability,

Audition processes in which abilities on musical perception and musical expression are evaluated by a jury commission.

Standardized Tests

Standardized tests are designed to measure the aural abilities in the perception of various musical elements. These exams are in multiple-choice test format in which the applicants are exposed to sound coming from speakers (or headphones) and are expected to answer questions regarding

abstracted musical elements - such as volume, dynamics, musical interval, timbre, texture, tempo, rhythm, melody and harmony – by making certain comparisons and discriminations. The most known examples of this method are Seashore test, Wing test, Bentley test and Gordon tests (including Gordon MAP, Gordon PMMA and Gordon IMMA). In his chapter on Musical Aptitude Tests, Tarman gives a detailed discussion of these test designs (Tarman, 2016:103-113). Similar designs have also been implemented in Turkey such as DYT ("Deneme Yetenek Testi" – Aptitude Trial Test) (Göğüş, 1994), MÖZYES ("Merkezi Özel Yetenek Sınavı" – Central Special Talent Exam) – which, according to Tarman was implemented only twice during the exams of 1994-1995 and 1995-1996 (Tarman, 2016: 113) -, OMÜ-MAT ("Ondokuz Mayıs Üniversitesi Müziksel Algılama Testi" – Ondokuz Mayıs University Musical Aptitude Test) (Ibid. 114), and MAÖ ("Müziksel Algılama Ölçeği" – Measure of Musical Perception) (Atak Yayla, 2009:372-377). There are two main advantages of this type of multiple choice tests; first of all, since each applicant is asked the same question, the same way and evaluated equally, their evaluation results are much more objective when compared to that of jury committee evaluations. Secondly, they use much less time and energy; an assigned exam superintendent can carry on the exam procedure in a room or a conference hall with as many applicants as possible at the same time, and the multiple choice answer sheets can be quickly processed later through an optical reader. Yet, besides these two important advantages, the usefulness of these test designs are also open to debate. The first problematic is their multiple-choice nature; some of these exams have questions that only provide two choices, thus the applicant has a 50% chance to score correct even if (s)he doesn't have any idea about the answer. The processed sounds that are played back during the exam and the acoustics of the exam space are other issues; some of these questions use unnatural sounds (such as an oscillator or a MIDI) which also result in an alienated nature from musicality, and the speaker system placed within the room/hall might cause individual differences in perception of sounds according to the acoustics of the space. However, probably the most important factor is that, although these types of exams may measure the individual aural abilities of a person to some degree, it is still a question whether these abstracted abilities correspond to a potential for musicality. (To give an example; from the results of their designed tests, Atak Yayla and Yayla (2009) investigated the predictive power of their test results with the musical talent of those who took the test. The results, although they were in positive correlation, showed a medium-low level relationship (r=0,483, r2 = 0,234))

That is why these tests – if they are used – are preferred more as a qualification exam in Turkey and have a filtering function; once an applicant passes these exams, (s)he is entitled to enter the final entrance exam, which is held by a jury committee.

Jury Committee Based Exams

Although the design of the jury exams - in which the applicant's musical talents are evaluated by assigned jury committees - vary according to the respective institutions preferences; they are usually evaluated within two main criteria: pitch recognition (including single pitches, intervals and chords), musical memory (both melodic and rhythmic). In each of these, the candidate is required to sing or play back what has been played for her with the piano reference. There can also be additional questions such as melodic/rhythmic dictation and/or melodic/rhythmic sight singing, however as these questions also require a musical knowledge besides talent, they usually are not encountered in the qualification (first) exams that fulfill a filtering function (if the entrance exams have two-tiers). Jury-based examination systems are much more preferred in Turkey than the standardized multiple-

choice tests. A nation-wide survey among Fine Art High School's music department teachers that was carried out by Yağcı (Yağcı, 2010:228) during the 2006-2007 education year showed that 9.2% of the surveyors totally agreed with the effectiveness of the jury based system, while 44.6% were in agreement to a large extent and 40% partially agreed. The rest, 6.2% thought that the effectiveness was very little. Thus, one can say that most of the teachers nationwide believed in the efficiency of this system. Nevertheless, held on a limited time with numerous applicants, these jury-based exams also bear many difficulties as they require the evaluation of each candidate separately. To give an example, in the 2015 musical entrance exams of ITU Turkish Music State Conservatoire, 5 different jury commissions, each consisting of 3 people, separately evaluated 507 candidates in 3 full days. As can be seen the amount of human resource, as well as time and energy devoted to this process is significantly high. Some of the shortcomings of this exam type is also related with this aspect, since a person may not be able to keep the same efficiency throughout such a long and tiring process. There is also the possibility of different jury committees developing different criteria for assessment during the exam period; that their reference performances (exam questions) may show differences (in terms of volume, tempo and accentuation); that the jury members may influence each other. In addition, the usage of limited number of manually created question packages in some cases (to avoid any leaking of the exam outside) may produce doubts about the equality of the difficulty level of the exam among all applicants. Such potential obstacles to an objective and a standardized measurement are the main disadvantages and the drawbacks of this system. Testing of jury reliabilities from the jury score sheets at first seems to offer a control mechanism (as seen in Atılgan (2008), Ece & Kaplan (2008), Tarman (2016:90)…etc.), yet, as Tarman also underlines, a high reliability score does not necessarily mean that the jury member had acted independently and/or evaluated objectively or consistently (Tarman, 2016:118)

(Surely one can avoid such pitfalls by some improvements such as increasing the number of jury members in a committee, isolating each jury member from each other - so that they would not know the scores of other members -, allowing longer time intervals for the jury to rest in between sessions…etc. A similar improved system is used in the music entrance exams of Yıldız University Department of Music and Performing Arts since the educational year of 2016-2017. Here, except for the head of the jury committee, each jury member is isolated from each other, and enter their scores to a computer they use individually. When the scoring of the applicant is finished, the head of the jury committee checks the variances between the jury members, and if there are huge differences ask them to reconsider scoring by playing the recorded version of the applicant's performance. Yet, as it is clear from this example, any of such improvements already result with additional costs.)

. In order to check those facts, one also needs to analyze the applicant's performance recordings through music/sound technologies and compare them with the jury scores. Thus, at this point the usage of Music Information Retrieval (MIR) technologies, which offers many approaches for automatic analysis of recorded sounds, might be a solution to overcome such disadvantages. The second part of the paper will present the findings of such a research project which investigated the effective potentiality of using sound engineering tools in the musical hearing portion of the musical aptitude exams.

Research Findings Concerning the Standardness of the Jury Based Exams

This part will present two important findings of a two-year research project (May 2016 – May 2018) that investigated the potential of using sound engineering tools in the musical hearing portion of the

musical aptitude exams. In general, the project tested the success of using such technological tools in evaluating the recorded sounds of the candidates by comparing the jury evaluations with computational analyses of the candidates' exam performance recordings. The jury evaluation reports (of the qualification exams of years 2015, 2016 and 2017) and the exam recordings (of years 2015 and 2017) were provided by Istanbul Technical University Turkish Music State Conservatory Music Theory department with the permission of the conservatory directorate. As the main goal was to make the qualification exams as efficient as possible, the project team also diagnosed some previously unobserved flaws about the question packages and offered some improvements for the exam preparation committee. Besides this, the most noticeable finding was that although the individual reliability scores of the jury committees were high (based on the jury reports), our computational analyses showed that each jury committee were developing different criteria especially when evaluating melodic memory sections; which brings to mind Tarman's doubts about the independency of the jury members in a jury committee (Ibid). Below we will be sharing these two main findings that may compromise the standardness of the jury based exams.

Problems about Different Question Packages

As it was stated earlier, nearly all jury-based exams in Turkey share two main criteria: pitch recognition (including single pitches, intervals, triads) and musical memory (melodic and rhythmic), although there also might be some extensions (sight singing, dictation or musical performance). Due to a high number of applicants, some of these institutions prefer a two-tier entrance exam, in which the first exam tests solely the previously mentioned musical abilities and functions more as a qualification for the final entrance exam. Thus the first (qualification) exam, although it takes less time for each applicant, is a long process that is conducted by different jury committees working simultaneously within multiple days. Such a setting requires additional precautions regarding the confidentiality of the questions asked in the exam. One of these precautions is designing the exam with various question packages; each package having its own set of distinct questions about pitch recognition, melody and rhythm - thus minimizing the chance of a leakage of the questions outside (i.e. memorization of a melody by a more specialized applicant and singing it back outside to her friends that are waiting for their turn). Yet, such a precaution may also create other problems, such as differences between the question packages in terms of their difficulty level. It is important to note that, the qualifications from these exams are not determined according to a ranking system exam, the applicants should score at least above a certain percentage; so the exam preparation committee takes this percentage of success into consideration not the ranking, and prepares the questions accordingly. Thus the applicants are expected to be successful above such a predetermined score regardless of which question package is used. However, even a mild variation between two question packages may produce amplified and significant differences in applicant performances due to unpredictable factors (applicant background, exam anxiety and individual capabilities…etc.). Bringing the exam closer to an ideally standard level starts from the equal distribution of question difficulties amoung various question packages.

The number of applicants we had analyzed the jury evaluations are; 365 people from the qualification exam of 2015, 456 people from 2016 and 451 people from 2017. The reliability scores of the jury

committees were in general very high as can be seen from Table 1, which will be discussed in the next section. The contents of these exams are as follows;

Pitch Recognition

Single Pitch Recognition (x5)

Interval Recognition (x5)

Triads (x4)

Musical Memory

Melodic Memory (Tonal & Modal; one question for each)

Rhythmic Memory (Straight & Aksak; one question for each)

2015 (n=365)	Jury Committee #1 (97 applicants)			Jury Committee #2 (100 applicants)			Jury Committee #3 (90 applicants)			Jury Committee #4 (78 applicants)		
	Avg. Pairwise Percent Agr.	Fleiss Kappa	K-Alpha	Avg. Pairwise Percent Agr.	Fleiss Kappa	K-Alpha	Avg. Pairwise Percent Agr.	Fleiss Kappa	K-Alpha	Avg. Pairwise Percent Agr.	Fleiss Kappa	K-Alpha
Single Pitch	98,76	0,798	0,798	98,13	0,888	0,888	100,00	1,000	1,000	99,32	0,830	0,830
Interval	95,60	0,894	0,894	93,47	0,867	0,867	97,04	0,929	0,929	95,56	0,903	0,903
Triad	95,88	0,906	0,906	92,67	0,852	0,853	97,22	0,939	0,939	93,59	0,857	0,857
Melody 1 (Tonal)	92,78	0,884	0,885	95,33	0,928	0,928	97,04	0,956	0,956	93,16	0,904	0,904
Melody 2 (Modal)	89,35	0,849	0,850	91,33	0,876	0,877	94,07	0,916	0,916	84,62	0,780	0,781
Rhythm 1 (Straight)	91,75	0,870	0,871	93,00	0,900	0,900	94,07	0,918	0,919	89,74	0,828	0,829
Rhythm 1 (Aksak)	86,94	0,779	0,780	95,33	0,928	0,928	99,26	0,985	0,986	95,73	0,913	0,914
TOTAL EXAM	95,34	0,910	0,910	94,57	0,910	0,910	97,70	0,955	0,955	95,11	0,906	0,906

2016 (n=456)	Jury Committee #1 (117 applicants)			Jury Committee #2 (115 applicants)			Jury Committee #3 (111 applicants)			Jury Committee #4 (113 applicants)		
	Avg. Pairwise Percent Agr.	Fleiss Kappa	K-Alpha	Avg. Pairwise Percent Agr.	Fleiss Kappa	K-Alpha	Avg. Pairwise Percent Agr.	Fleiss Kappa	K-Alpha	Avg. Pairwise Percent Agr.	Fleiss Kappa	K-Alpha
Single Pitch	99,66	0,944	0,944	98,26	0,873	0,873	99,76	0,973	0,973	98,70	0,938	0,938
Interval	97,49	0,944	0,944	98,26	0,873	0,873	97,84	0,952	0,952	95,40	0,908	0,908
Triad	96,72	0,933	0,933	92,90	0,858	0,858	96,85	0,936	0,936	94,69	0,893	0,893
Melody 1 (Tonal)	97,15	0,962	0,962	83,77	0,776	0,777	99,40	0,991	0,991	88,79	0,846	0,846
Melody 2 (Modal)	93,73	0,900	0,900	88,70	0,816	0,817	97,60	0,962	0,962	85,25	0,749	0,750
Rhythm 1 (Straight)	93,16	0,906	0,907	79,42	0,724	0,725	97,60	0,968	0,968	80,83	0,737	0,738
Rhythm 1 (Aksak)	93,16	0,903	0,903	80,29	0,722	0,723	97,00	0,958	0,958	83,48	0,746	0,747
TOTAL EXAM	97,21	0,946	0,946	92,69	0,875	0,875	98,17	0,966	0,966	93,76	0,897	0,897

2017 (n=451)	Jury Committee #1 (163 applicants)			Jury Committee #2 (159 applicants)			Jury Committee #3 (131 applicants)		
	Avg. Pairwise Percent Agr.	Fleiss Kappa	K-Alpha	Avg. Pairwise Percent Agr.	Fleiss Kappa	K-Alpha	Avg. Pairwise Percent Agr.	Fleiss Kappa	K-Alpha
Single Pitch	99,51	0,966	0,970	98,91	0,915	0,915	99,59	0,951	0,951
Interval	93,54	0,865	0,870	96,90	0,935	0,935	98,05	0,960	0,960
Triad	91,51	0,830	0,830	98,11	0,962	0,962	97,69	0,954	0,954
Melody 1 (Tonal)	87,73	0,826	0,830	93,08	0,905	0,906	97,44	0,965	0,965
Melody 2 (Modal)	91,00	0,772	0,770	89,10	0,754	0,754	95,90	0,924	0,924
Rhythm 1 (Straight)	85,28	0,798	0,800	87,21	0,827	0,827	92,56	0,899	0,900
Rhythm 1 (Aksak)	85,89	0,801	0,800	87,63	0,832	0,833	93,59	0,913	0,913
TOTAL EXAM	93,40	0,909	0,910	96,03	0,946	0,946	97,69	0,968	0,968

Table 1: 2015-2017 Analized Reports: Jury Reliability Scores Using Various Measurement Tests

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
SinglePitch_Success	Between Groups	,272	9	,030	1,303	,234
	Within Groups	8,225	355	,023		
	Total	8,497	364			
Interval_Success	Between Groups	1,963	9	,218	2,084	,030
	Within Groups	37,157	355	,105		
	Total	39,120	364			
Triad_Success	Between Groups	2,225	9	,247	2,246	,019
	Within Groups	39,083	355	,110		
	Total	41,308	364			
Melody1_Success	Between Groups	6,951	9	,772	5,982	,000
	Within Groups	45,832	355	,129		
	Total	52,783	364			
Melody2_Success	Between Groups	8,501	9	,945	7,263	,000
	Within Groups	46,166	355	,130		
	Total	54,667	364			
Rhythm1Success	Between Groups	7,564	9	,840	8,687	,000
	Within Groups	34,345	355	,097		
	Total	41,909	364			
Rhythm2Success	Between Groups	6,170	9	,686	6,309	,000
	Within Groups	38,578	355	,109		
	Total	44,749	364			
Total_Success	Between Groups	12088,975	9	1343,219	2,629	,006
	Within Groups	181380,569	355	510,931		
	Total	193469,545	364			

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
SinglePitch_Success	Between Groups	,852	9	,095	2,339	,014
	Within Groups	18,050	446	,040		
	Total	18,902	455			
Interval_Success	Between Groups	1,129	9	,125	1,091	,368
	Within Groups	51,278	446	,115		
	Total	52,407	455			
Triad_Success	Between Groups	1,873	9	,208	1,912	,048
	Within Groups	48,547	446	,109		
	Total	50,421	455			
Melody1_Success	Between Groups	2,974	9	,330	2,253	,018
	Within Groups	65,406	446	,147		
	Total	68,381	455			
Melody2_Success	Between Groups	3,240	9	,360	3,260	,001
	Within Groups	49,255	446	,110		
	Total	52,495	455			
Rhythm1Success	Between Groups	10,103	9	1,123	9,024	,000
	Within Groups	55,483	446	,124		
	Total	65,587	455			
Rhythm2Success	Between Groups	10,037	9	1,115	7,940	,000
	Within Groups	62,644	446	,140		
	Total	72,681	455			
Total_Success	Between Groups	12248,682	9	1360,965	2,311	,015
	Within Groups	262677,646	446	588,963		
	Total	274926,328	455			

Table 2: 2015 & 2016 Exams ANOVA Tests – Success vs. Question Packages

Table 2 presents the ANOVA results obtained from the 2015 and 2016 tests, considering the possible effect of using 10 different question packages on the success of the applicants. Generally speaking, both the F values and the *p* values suggest that, for each category the mean success percentage is significantly different for at least one of the question packages. Especially the melodic and rhythmic memory categories were the most problematic in this sense. Thus, considering the "Total_Success" category, which is the exam score of the applicants, one can conclude that the test score of an applicant was also dependent on which question package she was evaluated according to. However, this surely doesn't mean the dependency of passing/failing the exam to the question packages. Table 3 presents the same effect on the exam qualifications (for the 2015 exam the qualification score was 60%, for the 2016 exam it was 50%). We observe that, both in 2015 and 2016, there wasn't any significant relationship between the passing/failing of an applicant with her assigned question package ($p > 0{,}05$ for both).

ANOVA

Pass_2015

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3,711	9	,412	1,799	,067
Within Groups	81,357	355	,229		
Total	85,068	364			

ANOVA

Pass_2016

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3,835	9	,426	1,776	,071
Within Groups	106,998	446	,240		
Total	110,833	455			

Table 3: 2015 & 2016 Exams ANOVA Tests – Pass/Fail vs. Question Packages

With these information and data our research team designed and conducted an experiment following the exam of 2016. The experiment was modeled from the questions of 2016 exam, and its aim was to test the degree of variation between the question packages among the music conservatory students – these are those people we assume the question package choice does not play a role in the success of the candidate. The information that would be obtained from this research, in addition to the previous data, would not only help us understand the difficulty levels and the ease of perception of the questions but may also suggest improvements for our question designs. We have conducted the experiment with 26 students from Musicology and Music Theory departments. The questions that were used were from three question packages used in 2016 qualification exams; those having the average (assigned as package#1), lowest (#2) and highest (#3) amount of successes from each category – thus the experiment was also checking the results of 2016 qualification exams. The MAST experiment (Musical Aptitude Standard Test) was conducted in the Musicology lab individually with

usage of computers, headphones and microphones. Similar to a TOEFL exam, the participants were asked to follow the instructions appearing on the screen in front of them. The questions were played from MIDI formats and the participants were asked to sing/perform what they heard on the headphones to the microphone on their desks. Meanwhile, one of the researchers of the project was recording the responses of the participants on a different computer. Thus there was no jury committee present in the room; the research team later compiled the recordings and sent them to a jury committee for evaluating. After the experiment, the participants also filled out a survey regarding the relative efficiency of this exam system when compared to a live jury committee system. Out of 26 people 6 preferred the jury system (23%), 6 were indifferent between the two systems (23%), while 14 people (54%) found this system better than the jury based system and wrote that such an environment had a positive effect on their efficiency. We should also note that the Musicology Lab in which the experiment was conducted had a poor sound isolation, and that the 6 participants who preferred the jury system also wrote in their comments that they were confused due to noise coming from outside to room if not they felt strange in such an isolated exam environment. The whole MAST experiment process, including the introduction, the experiment and the survey took around 15 to 20 minutes for each participant.

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
SinglePitch_Success	Between Groups	,002	2	,001	,037	,964
	Within Groups	2,194	66	,033		
	Total	2,196	68			
Interval_Success	Between Groups	,113	2	,056	,568	,570
	Within Groups	6,560	66	,099		
	Total	6,673	68			
Triad_Success	Between Groups	,031	2	,015	,189	,829
	Within Groups	5,368	66	,081		
	Total	5,399	68			
Melody1_Success	Between Groups	,507	2	,254	4,663	,013
	Within Groups	3,589	66	,054		
	Total	4,096	68			
Melody2_Success	Between Groups	,792	2	,396	6,936	,002
	Within Groups	3,766	66	,057		
	Total	4,558	68			
Rhythm1_Success	Between Groups	1,149	2	,575	11,596	,000
	Within Groups	3,270	66	,050		
	Total	4,419	68			
Rhythm2_Success	Between Groups	1,399	2	,699	13,425	,000
	Within Groups	3,439	66	,052		
	Total	4,837	68			
TOTAL_Success	Between Groups	,252	2	,126	4,322	,017
	Within Groups	1,920	66	,029		
	Total	2,172	68			

Table 4: 2016 MAST Experiment ANOVA Test – Success vs. Question Packages

Table 4 presents the ANOVA test results of the MAST experiment that was conducted using only three packages from the qualification exams of 2016. What is noticeable is that in the pitch recognition part (single pitch, interval and triad identification), there is no relationship between the assigned question packages and the degree of success. In other words, the level of difficulty of 2016 question packages were designed for the qualifiers based on our assumption that our experiment participants – already being conservatory students - are potential qualifiers of the exam. On the other hand, the results of the melodic and rhythmic memory sections came out parallel with that of 2016 qualification exam results. As can be seen from Figure 1 & Figure 2, the questions of package #2 – which were selected from the question packages with the lowest amount of success in 2016 - in all four categories (melodies 1&2, rhythms 1&2) got the lowest score as we well.

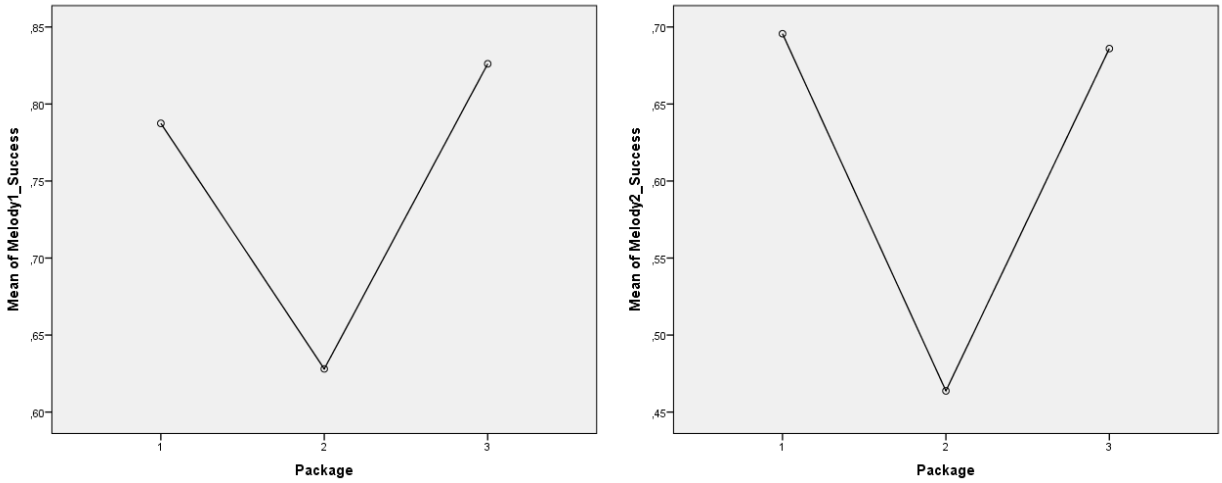

Figure 1: 2016 MAST Experiment: Means Plots for Melody Questions vs. Question Packages

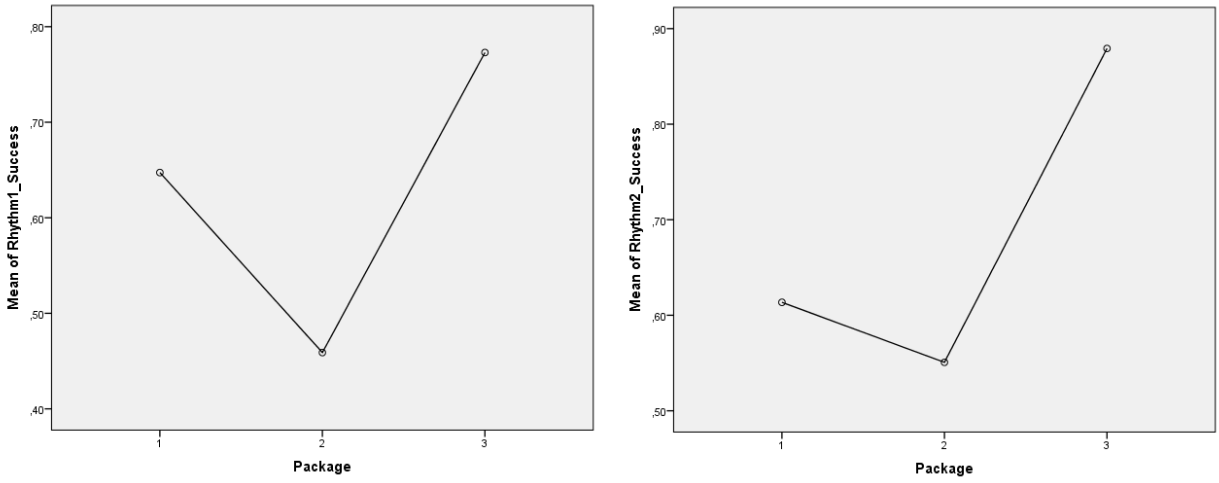

Figure 2: 2016 MAST Experiment: Means Plots for Rhythm Questions vs. Question Packages

As these results confirmed the results of the 2016 melody and rhythm question packages, our research team analyzed the possible factors that may cause such a difference in terms of the level of

difficulty. We realized that, although all the melody questions were two measures long, and all the rhythm questions one measure long; factors such as number of notes, range/ambitus, shape of the melody, the proportion of melodic steps with melodic leaps, the periodicity and the familiarity of the passage may also be contributing to this different levels of difficulties. Thus, we designed a rubric for preparing melody and rhythm questions and proposed it to the exam preparation committee before the preparation of 2017 qualification exams. The guidelines we had proposed were as follows;

Melody Questions

Each melody group should use the same number of notes,

The same note should not be used consecutively,

Each melody group should have the same rhythmic values,

Each melody group should have the same range/ambitus,

Each melody group should have the same time signature,

Each melody should be two measures long,

Each melody group should have the same tonality,

The melodies in each group should start and end with the tonic,

(For tonal melodies) The melodies should imply a similar harmonic background (such as I – ii – V7 – I),

(For modal melodies) The melodies should have a similar modal progression.

Rhythm Questions

Each rhythm should use the same rhythmic motifs arranged in different orders (like a-b-c-d; b-a-c-d; c-a-b-d; d-a-b-c; a-c-b-d …etc.),

The rhythms should not contain or imply a periodic structure (like a-b-a-c),

Each rhythm group should have the same time signature,

Each rhythm should be two measures long.

Besides these we also suggested the interval and triad questions be designed in such a way that not only their content but their order be arranged in similar ways. Table 5 presents the results of the 2017 qualification exam regarding the relationship between success and question packages.

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
SinglePitch_Success	Between Groups	6538,904	9	726,545	1,878	,053
	Within Groups	170578,471	441	386,799		
	Total	177117,374	450			
Interval_Success	Between Groups	16947,234	9	1883,026	1,647	,100
	Within Groups	504238,789	441	1143,399		
	Total	521186,024	450			
Triad_Success	Between Groups	7476,920	9	830,769	,708	,702
	Within Groups	517825,967	441	1174,209		
	Total	525302,887	450			
Melody1_Success	Between Groups	40962,279	9	4551,364	3,189	,001
	Within Groups	629456,649	441	1427,339		
	Total	670418,927	450			
Melody2_Success	Between Groups	24078,598	9	2675,400	4,127	,000
	Within Groups	285877,883	441	648,249		
	Total	309956,482	450			
Rhythm1_Success	Between Groups	76598,752	9	8510,972	6,260	,000
	Within Groups	599584,265	441	1359,602		
	Total	676183,018	450			
Rhythm2_Success	Between Groups	101074,461	9	11230,496	8,377	,000
	Within Groups	591210,977	441	1340,614		
	Total	692285,437	450			
Total_Success	Between Groups	6550,710	9	727,857	1,538	,132
	Within Groups	208643,085	441	473,114		
	Total	215193,794	450			

Table 5: 2017 Exam ANOVA Test – Success vs. Question Packages

The observable decrease in the F values and the increase in *p* values in the pitch recognition section (single pitches, intervals and triads), which now suggests no relationship between the question packages with the success in these categories demonstrates the benefits of using the same content in the same order – thus just transposed versions of the same question – in the interval and triad sections. It seems that the guidelines we had proposed earlier for the preparation of melody and rhythm questions did not have any positive effects on these sections though, as the amount of success in all four categories (melody1, melody2, rhythm1 and rhythm2) still show a dependency with at least one of the question packages (*p* =< 0.001 in each). From the Bonferroni post-hoc tests we spotted the means of one of the melodies from melody1 package, two of the melodies from melody2 package and more than two of the rhythms in the rhythm1 and rhythm2 packages were significantly different than the means of the corresponding questions of the other packages. However, for the first time, the total score gained from the exam resulted with an insignificant relationship; with the values F= 1,538 and *p* = 0,132. In addition to this, as can be observed from Table 6, there was no significant relationship on the passing/failing of the exam with the question package (for the 2017 exam the qualification score was 50%); with the values F= 0,951 and *p* = 0,480 - the lowest F and the highest *p* values throughout the research so far. The reason for this is probably the fact that the significantly different questions in terms of the level of difficulty were dispersed among different question packages for each category; ex. a package containing melody1 with the lowest mean, whereas the mean of the melody2 from the same package had a high mean. Thus, such results of the 2017 qualification exams may be an outcome of a coincidence.

ANOVA

Pass_2017

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2,108	9	,234	,951	,480
Within Groups	108,579	441	,246		
Total	110,687	450			

Table 6: 2017 Exam ANOVA Tests – Pass/Fail vs. Question Packages

It seems that more research is needed in the area of music perception for the standardization of such melody and rhythm questions. The possible flaws of using different question packages could be significantly minimized using such question preparation guidelines with generative rubrics. In addition to this, in order to increase the amount of dispersion, multiple questions – at least three - should be asked for each category rather than only one; ex. three tonal melody questions, three modal melody questions …etc. However, this will also result in an increase in the amount of time used per applicant. The solution we would propose will be discussed in the final section.

Reconsidering Jury Reliabilities

As the main aim of our research was to investigate the design and applicability of automatic assessment tools to support the qualifying exams, the recorded sounds of the candidates' exam performances were analyzed in comparison with the evaluations of the jury members. In other words, the algorithms were designed not to decide whether a reference piano sound matched with an applicant performance or not, but to imitate the jury responses and make an evaluation about the applicant performances. Here, for the pitch recognition section, we have managed to find reliable information on the acceptable pitch ranges and thresholds about the interval intonations (which are investigated and discussed separately in Köker et al. and Güner et al.). The hardest category to accomplish the automatic assessment task was the melody sections; since evaluation of a "successful" melodic recall may have multiple factors including, the completeness of the melody, pitch row, rhythm, intonation, melodic shape, vibrato …etc., and that the importance of these factors may change from a person to person. Using the dataset derived from 2015 and 2016 qualification exams, two different assessment systems have been designed (as discussed in Bozkurt et al. (2017) and Gültekin et al.), having average accuracies as 0.74 and 0.856. During this process we also had the chance to analyze the samples in which the automatic assessment tool and the jury evaluations had disagreements. Except for a few cases, it was observed that all the disagreements were on the ones in which the applicants were favored by the jury committee; that is the jury committees (of the 2015 and 2016 exams) turned out to be more positively flexible than the algorithm, evaluating the applicants as successful in cases that might be considered as unsuccessful. Such a separation also provided us with information about the different evaluation criteria of the different jury committees; intonation, rhythm, the place (and the function) of the missed note …etc. In fact, the existence of different criteria between jury committees was probably one of the causes for the automatic assessment systems – which "learn" how to evaluate from these different committees - not having higher accuracies.

Figure 3 presents the overall distribution of the 2016 & 2017 qualification exam scores. The base score needed for qualifying in these years were 50%. The bold marked bar in both histograms refers to the area of 48%-51,9%. From the frequency distributions we observe that most of the applicants that fall in this area actually passed the exam (19/24 passed in 2016 and 20/22 passed in 2017 as can be seen from Table 7). Also notice from the two histograms how drastic the differences are with the bars on their left (fail) and the right (pass). It is as if a "positive" transfer has been made from the left to the right, which, for each jury member, requires only 5 points - that is the difference between a "totally successful melody" with an "averagely successful melody" provided in the jury evaluation sheets. Thus, it seems that the jury committee is acting as a group, and by taking the initiative altogether, deciding to give a second chance for the applicant in the final entrance exams. This "jury-induced" positive effect might also explain those cases in which the jury score and the designed algorithm were in disagreement we have stated before.

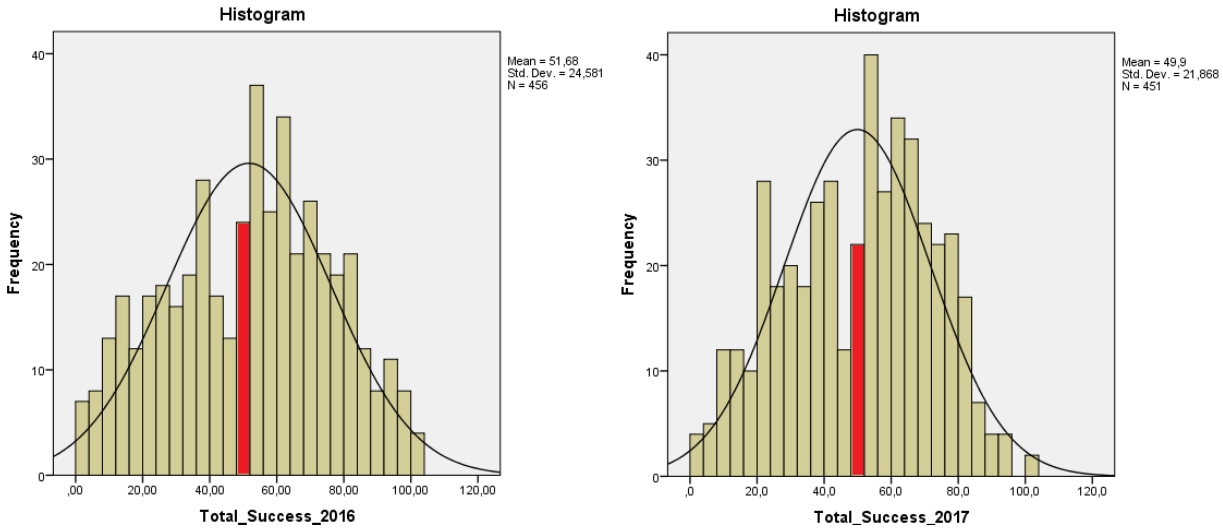

Figure 3: 2016 & 2017 Qualification Exams – Overall Distribution of Scores

n = 24
Fail = 5
Pass = 19

Total_Success_2016

	Frequency	Percent	Valid Percent	Cumulative Percent
44,00	4	,9	,9	38,6
44,70	1	,2	,2	38,8
45,00	5	1,1	1,1	39,9
46,00	1	,2	,2	40,1
46,70	2	,4	,4	40,6
48,30	1	,2	,2	40,8
49,00	2	,4	,4	41,2
49,30	1	,2	,2	41,4
49,70	1	,2	,2	41,7
50,00	8	1,8	1,8	43,4
50,30	2	,4	,4	43,9
50,70	3	,7	,7	44,5
51,00	2	,4	,4	45,0
51,30	2	,4	,4	45,4
51,70	2	,4	,4	45,8
52,00	2	,4	,4	46,3
52,30	2	,4	,4	46,7
53,00	14	3,1	3,1	49,8
53,30	3	,7	,7	50,4
54,00	4	,9	,9	51,3

n = 22
Fail = 2
Pass =20

Total_Success_2017

	Frequency	Percent	Valid Percent	Cumulative Percent
44,0	3	,7	,7	40,8
44,7	2	,4	,4	41,2
45,0	1	,2	,2	41,5
45,3	2	,4	,4	41,9
46,7	1	,2	,2	42,1
47,0	3	,7	,7	42,8
48,0	1	,2	,2	43,0
48,3	1	,2	,2	43,2
50,0	7	1,6	1,6	44,8
50,3	1	,2	,2	45,0
50,7	1	,2	,2	45,2
51,0	2	,4	,4	45,7
51,3	4	,9	,9	46,6
51,7	5	1,1	1,1	47,7
52,0	3	,7	,7	48,3
52,3	2	,4	,4	48,8
52,7	1	,2	,2	49,0
53,0	10	2,2	2,2	51,2
53,3	3	,7	,7	51,9
53,7	2	,4	,4	52,3
54,0	7	1,6	1,6	53,9

Table 7: 2016 & 2017 Qualification Exams: 48%-51,9% Score Area

At first this may not seem as a negative thing, especially when considered from the perspective of the applicants. However, it also shows that a communication between the jury members is present, which may occur in other cases as well, and thus be positively contributing to the high reliability scores we have presented in the previous section. In addition, bear in mind that, in Turkey, those applicants who are graduated from the Fine-Arts High Schools (*Güzel Sanatlar Lisesi*), gain a significant amount of extra points than ordinary high school graduates at the final entrance exams. Such applicants might appear at the top portion of the conservatory acceptance list, even if they had ended up actually in the waiting lists as a result of their final entrance exam performances. Thus it is open to discussion whether elevating a failing applicant coming from a Fine Arts High school background above the base score is a "positive" act, especially when considered from the perspective of those coming from ordinary school backgrounds who have passed the qualification exams without any outside effects.

Conclusion: Towards a Hybrid Assessment Model

As a conclusion we propose a qualification exam design that is similar to the MAST experiment that was mentioned previously. Thus, similar to the TOEFL or PEARSON English exams, the applicants would register and individually have their qualification examination in isolated booths with the usage of computer screens, headphones and microphones through which their questions would be asked and their live responses would be recorded, assessed using state of the art MIR technologies and further screened by instructors especially for the close to boundary cases. Recall from the survey results of the MAST experiment that such a new system was preferred by the majority of the participants, and that the 23% who preferred the jury system based their complaints on the poor isolation of the exam environment – which can be prevented using a better room. Considering the question packages we have discussed in the previous section, the system could ask different questions for each category from a randomly selected big pool of different packages; thus generating a different exam each time. It is also possible to create such questions by computers through generative algorithms (Currently, a similar algorith is in design and testing phase). The test can also incorporate similar type of hearing-based multiple choice questions as seen in the standardized tests of Seashore, Wing, Bentley and Gordon. In such an exam system the applicants could also get a detailed evaluation report of their exam performances, and - if they have scored above a base score - use these reports to apply for the final entrance exams for the music conservatories, in which they can show their musical performance skills to the jury committee in more detail. This will not only save a great deal of time, energy, infrastructure and capital, but will also increase the quality of the final entrance exams (the second tier exams), and result in a much more efficient examination process.

Acknowledgements: This work is supported by the Scientific and Technological Research Council of Turkey, TUBITAK, Grant#215K017.

References

Atak Yayla, A., Yayla, F. 2009. "Müziksel Algılama Ölçeği". *8. Ulusal Müzik Eğitimi Sempozyumu: Türkiye'de Müzik Eğitiminin Sorunları ve Çözüm Önerileri – Bildiriler Kitabı.* Ondokuz Mayıs Üniversitesi Yayınları. Samsun. 372-378.

Atılgan, H. 2008. “Using Generalizability theory to assess the score reliability of the Special Ability Selection Examinations for music education programmes in higher education”. *International Journal of Research & Method in Education*, 31:1, 63-76.

Bozkurt, B., Baysal, O., Yüret, D. (2017). “A Dataset and Baseline System for Singing Voice Assessment”. CMMR 2017 13th International Symposium on Computer Music Multidisciplinary Research: Music Technology with Swing. 25-28 September 2017.

Ece, A. S., Kaplan, S. 2008 .“Müzik Özel Yetenek Seçme Sınavı’nın Puanlayıcılar Arası Güvenilirlik Çalışması”. National Education, 36-49.

Göğüş, G. 1999. “Müzik Yeteneğinin Tanımı, Ölçümü ve Deneme Yetenek Testi”, *Uludağ Üniversitesi Eğitim Fakültesi Dergisi*, Cilt: 12, sayı: 1. 79-89.

Güner, B.B., Baysal, O., Bozkurt, B. (in preparation). “Müzik Yetenek Sınavları Çift Ses Soru Değerlendirmelerinde Kabul Edilebilir Aralıklar”.

Gültekin, C., Bozkurt, B., Baysal, O. (in preparation). “Singing Assessment Using Chroma Features”.

Köker, O., Baysal, O., Bozkurt, B.(forthcoming). “Müzik Yetenek Sınavlarında Tek Ses Tekrarları İçin Kabul Edilebilir Perde Aralığı (Aralıkları)”. *Hacettepe Üniversitesi Ankara Devlet Konservatuarı Ulusal Müzik ve Sahne Sanatları II. Sempozyumu - Bildiri Kitabı* – 21 Aralık 2017, Ankara.

Tarman, S. 2016 (2006). *Müzik Eğitiminin Temelleri* – Geliştirilmiş 2. Basım. Müzik Eğitimi Yayınları. Ankara.

Yağcı, U. 2010. “AGSL Müzik Bölümleri Yetenek Sınavları ve Bu Sınavlara Yönelik Öğretmen Görüşleri”. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi*, Sayı 27, 223-231.