Exploring the Music Perception Skills of Crowd Workers

Samiotis, I.P.; Qiu, S.; Lofi, C.; Yang, J.; Gadiraju, Ujwal; Bozzon, Alessandro

Publication date
2021

Document Version
Final published version

Published in
Proceedings of the AAAI Conference on Human Computation and Crowdsourcing

Citation (APA)
Samiotis, I. P., Qiu, S., Lofi, C., Yang, J., Gadiraju, U., & Bozzon, A. (2021). Exploring the Music Perception Skills of Crowd Workers. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, *9*, 108-119. https://ojs.aaai.org/index.php/HCOMP/article/view/18944

Proceedings of the Ninth AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2021)

Exploring the Music Perception Skills of Crowd Workers

Ioannis Petros Samiotis, Sihang Qiu, Christoph Lofi, Jie Yang, Ujwal Gadiraju, Alessandro Bozzon
Delft University of Technology, Delft, Netherlands
{i.p.samiotis, s.qiu-1, c.lofi, j.yang-3, u.k.gadiraju, a.bozzon}@tudelft.nl

Abstract

Music content annotation campaigns are common on paid crowdsourcing platforms. Crowd workers are expected to annotate complicated music artefacts, which can demand certain skills and expertise. Traditional methods of participant selection are not designed to capture these kind of domain-specific skills and expertise, and often domain-specific questions fall under the general demographics category. Despite the popularity of such tasks, there is a general lack of deeper understanding of the distribution of musical properties - especially auditory perception skills - among workers. To address this knowledge gap, we conducted a user study ($N = 100$) on Prolific. We asked workers to indicate their musical sophistication through a questionnaire and assessed their music perception skills through an audio-based skill test. The goal of this work is to better understand the extent to which crowd workers possess higher perceptions skills, beyond their own musical education level and self reported abilities. Our study shows that untrained crowd workers can possess high perception skills on the music elements of *melody*, *tuning*, *accent* and *tempo*; skills that can be useful in a plethora of annotation tasks in the music domain.

Introduction

Several studies have shown the ability of crowd workers to successfully contribute to the analysis and annotation of multimedia content, both based on simple perceptual skills (e.g. for image analysis (Sorokin and Forsyth 2008)) and domain-specific knowledge (Oosterman et al. 2015). Musical content is no exception, and research has shown that the general crowd can be successfully involved in the annotation (Samiotis et al. 2020) and evaluation (Urbano et al. 2010) processes of music-related data and methods. Plenty of music annotation tasks (Lee 2010; Lee and Hu 2012; Mandel, Eck, and Bengio 2010; Lee, Hill, and Work 2012; Speck et al. 2011) can be routinely found on microtask crowdsourcing platforms, mostly focused on descriptive (Law et al. 2007) and emotional (Lee 2010) tagging.

Music, as a form of art, often requires a multifaceted set of skills to perform, and certain expertise to analyse its artefacts. There are cases that require advanced music perceptual skills (such as the ability to perceive changes in melody) and music-specific knowledge. However, both in literature and in practice, it is rare to encounter such crowdsourcing tasks. Consider, for example, annotation tasks targeting classical music, e.g. music transcription, performance evaluation, or performance annotation. Classical music is a genre featuring artworks with high musical complexity; it is no surprise that corresponding analysis and annotation tasks are often exclusively performed by musical experts and scholars. This unfortunately hampers current efforts to digitize and open up classical music archives, as scholars and experts are expensive and not easily available. Here, the ability to utilise microtask crowdsourcing as an annotation and analysis approach could bring obvious advantages. But how likely it is to find advanced music-related perceptual skills on crowdsourcing platforms? With the goal of answering this broad research question, in this paper we scope our investigation on the following two aspects:

- **[RQ1]** To what extent are higher perceptual skills of melody, tuning, accent and tempo, present on microtask crowdsourcing platforms?
- **[RQ2]** How are different perception skills and self-reported music-related knowledge distributed among crowd workers?

Studies on human cognition and psychology, have shown that people can possess innate music perception skills, without previous formal training (Mankel and Bidelman 2018; Ullén et al. 2014). However, the majority of those studies have been conducted in labs, under controlled conditions, and with limited amounts of participants.

In our work we set out to measure the music sophistication and perception skills of crowd workers operating on the Prolific crowdsourcing platform.[1]

We designed a rigorous study that employs validated tools to measure the musical sophistication of the users and quantify their music perception skills: the Goldsmith's Music Sophistication Index (GMSI) questionnaire (Müllensiefen et al. 2014) and the Profile of Music Perception Skills (PROMS) active skill test (Law and Zentner 2012) respectively. These tools allow for a general overview of musical ability characteristics, but also a more detailed understanding through their subcategories (e.g. musical training and melody per-

[1] https://www.prolific.co

ception skills). By juxtaposing passive methods of assessment (questionnaire) with the active evaluation of auditory skills, we aim to gather a better understanding of workers' actual skills on musical aspects, beyond their subjective self-assessment. With GMSI, we are able to evaluate a person's ability to engage with music through a series of questions focusing on different musical aspects. PROMS on the other hand, allows for a more objective way to measure a person's auditory music perception skills (e.g. melody, tuning, accent and tempo perception) through a series of audio comparison tests. To the best of our knowledge, this is the first attempt to use PROMS in an online crowdsourcing environment, and the measured perception skills can offer valuable insights to the auditory capabilities of the crowd.

Our findings indicate that pre-existing musical training is not common among crowd workers, and that music sophistication aspects are not necessarily predictive of actual music perception skills. Instead, we observe that the majority of workers show an affinity with specific sets of skills (e.g., we found a surprising number of *musical sleepers* — workers without formal training but still high music perception skill test results). As a whole, our study paves the way for further work in worker modelling and task assignment, to allow a wider and more refined set of microtask crowdsourcing tasks in the domain of music analysis and annotation.

Related Work

There is a long history of studies on perception and processing of music by humans; from the analysis of the socio-cultural variables influencing a person's musicality amplitude (Hannon and Trainor 2007), to the study of musicality from a genetics' base (Gingras et al. 2015). In all cases, inherent music processing capabilities have been found in people and they seem to be connected with basic cognitive and neural processes of language since early stages of development (Liberman and Mattingly 1985; Koelsch et al. 2009). Even people with *amusia*, a rare phenomenon where a person can't distinguish tonal differences between sounds (Peretz and Hyde 2003), they can still process and replicate rhythm correctly (Hyde and Peretz 2004).

In (Müllensiefen et al. 2014), we find a large scale study on musical sophistication through the use of the GMSI survey, on a unique sample of 147,663 people. GMSI is particularly calibrated to identify musicality in adults with varying levels of formal training. It is targeted towards the general public, and can prove less effective to distinguish fine differences between highly trained individuals. Musical sophistication in the context of that study, and ours, encompasses musical behaviours and practices that go beyond formal training on music theory and instrument performance. Their findings show that musical sophistication, melody memory and musical beat perception are related. The survey has been translated and replicated successfully (on smaller samples) in French(Degrave and Dedonder 2019), Portuguese(Lima et al. 2020), Mandarine(Lin et al. 2021), and German(Schaal, Bauer, and Müllensiefen 2014).

Our study draws connections to those findings and aims to shed light into the musical capabilities of people on crowdsourcing platforms. The demographics and conditions of the studies presented so far, cannot be easily compared to those of online markets. Users on those platforms are participating in such studies through monetary incentives, and the conditions (equipment, location, potential distractions, etc.) under which they perform the tasks cannot be controlled as in a lab environment (Totterdell and Niven 2014; Zhuang and Gadiraju 2019; Gadiraju et al. 2017b).

Currently, crowdsourced music annotation is primarily utilised for descriptive (Law et al. 2007) and emotional(Lee 2010) tagging. Large-scale music data creation and annotation projects such as Last.fm[2] and Musicbrainz[3], are largely depended on human annotation, but from users of their respective online social platforms. A survey on the applicability of music perception experiments on Amazon Mechanical Turk (Oh and Wang 2012), showed that online crowdsourcing platforms have been underused in the music domain and the status has not changed radically since then. Through our study, we want to examine the capabilities of the crowd on processing music audio and showcase their capabilities, in an attempt to encourage further research and utilisation of crowdsourcing in the music domain.

Experimental Design

The main focus of this study is to offer insights into the musical characteristics and perception skills of workers operating on crowdsourcing platforms. We therefore designed our experiment to capture these attributes through methods that can be used online, and that do not require pre-existing musical knowledge. We used two methods: 1) the *GMSI* questionnaire to evaluate the *musical sophistication* (musical training, active engagement and other related musical characteristics) (Müllensiefen et al. 2014)) of workers and 2) the *Mini-PROMS* test battery to evaluate their auditory music perception skills. We then compare the obtained results, paying specific attention to the overlapping aspects of musical sophistication and music perception skills. With this experiment, we are also interested in identifying "*musical sleepers*" and "*sleeping musicians*", a notion originally presented in (Law and Zentner 2012). A musical sleeper is a person with little to no musical training but with high performance in the perception test, while a sleeping musician indicates the opposite.

Procedure

After a preliminary step where workers are asked basic demographic information (age, education, and occupation), the study is composed of four consecutive steps (Figure 1), each devoted to collecting information about specific attributes corresponding to the crowd workers: 1) Musical Sophistication Assessment (*GSMI*), 2) Active Music Perception Skill Assessment (*Mini-PROMS*), 3) Self-Assessment of Music Perception Skills, and 4) Post-task Survey collecting information on workers audio-related conditions, and perceived cognitive load.

[2]https://www.last.fm

[3]https://musicbrainz.org

Figure 1: The five steps in the music perception skills study.

Questionnaires and Measures

Capturing Musical Sophistication of Workers. Musical behaviours of people such as listening to music, practicing an instrument, singing or investing on vinyl collections, all show the affinity of a person towards music. The degree to which a person is engaged to music through these behaviours, constitutes the musical sophistication. Musical sophistication can be measured as a psychometric construct through the *GMSI* questionnaire, which collects self-reported musicality through emotional responses, engagement with music, formal training, singing capabilities and self-assessed perception skills. It is an instrument specifically designed to capture the sophistication of musical behaviours, in contrast to other questionnaires such as Musical Engagement Questionnaire (MEQ) (Werner, Swope, and Heide 2006), which measures the spectrum of psychologial facets of musical experiences. More specifically, the musical sophistication of people based on (Müllensiefen et al. 2014), is organised into the following five facets:

Active Engagement: this aspect determines the degree to which a person engages with music, by listening to and allocating their time/budget to it;

Perceptual Abilities: this aspect assesses the skill of perceiving (mainly auditory) elements of music. This is an important subscale in our study, since the self-assessed perceptual skills of the workers in GMSI can be directly compared to those we actively measure in Mini-PROMS;

Musical Training: this aspect reports the years of training on aspects of music (e.g. theory, performing an instrument), which can indicate the formal expertise that a person has in the domain;

Emotions: this aspect determines the emotional impact of music on that person;

Singing Abilities: this aspect evaluates the ability to follow along melodies and tempo (beat) of songs.

GMSI offers additional questions outside the subscales, which capture specific properties of the participant: 1) "Best Instrument", which represents which instrument the user knows to play the best, 2) "Start Age", which age the participant starting learning an instrument and 3) "Absolute Pitch", which indicates if the person can understand correctly the exact notes of a sound frequency. Absolute pitch is a very rare trait that develops during the early stages of auditory processing (Burkhard, Elmer, and Jäncke 2019) but can deteriorate through the years (Baharloo et al. 1998). As such, a person with perfect pitch perception, could have an advantage on a melody perception test, thus we included it with the rest of the subscales.

The original GMSI questionnaire contains 38 main items and 3 special questions, and considering the rest of the study's parts, we chose to reduce its size, while keeping its psychometric reliability. For that purpose, we consulted the GSMI online "configurator"[4] which allows to select the number of items per subscales and estimates the reliability of the resulting questionnaire based on the questions it selects. We reduced the size of the questionnaire to 34 questions, and preserved the special question about "Absolute Pitch", resulting in 35 questions in total. Table 1 presents the psychometric values[5] of the final GMSI version we used; each sub-scale fares *very-good* to *excellent* reliability values (Hulin, Netemeyer, and Cudeck 2001).

In the GMSI questionnaire, each question from the subscales, uses the seven-point Likert scale (Joshi et al. 2015) for the user's responses, with most questions having "Completely Agree", "Strongly Agree", "Agree", "Neither Agree Nor Disagree", "Disagree", "Strongly Disagree" and "Completely Disagree" as options. Few questions offer numerical options for topics (e.g. indicating the time spent actively listening to music, or practicing an instrument). The workers is not aware of the subscale each question belongs to. The index of each subscale of GMSI is calculated with the aggregated results of the relevant questions. The overall index of "General Music Sophistication" is calculated based on 18 questions out of the total 34 items of the subscales; these 18 questions are predefined by the designers of the questionnaire; the question about "Absolute Pitch" does not contribute to the total index.

Using the GMSI questionnaire is close to the typical methods used to assess the knowledge background of annotators in other domains. Especially the questions of "Musical Training" follow standard patterns to assess the formal training of a person in a domain, thus a certain objectivity can be expected (assuming good faith from the workers). However, the rest of the categories are based purely on subjective indicators and self-reported competence, which can potentially misrepresent the true music behaviours and capabilities of a worker. For this reason, it is necessary to understand the best practices that could reliably predict a worker's performance to a music annotation task. To that end, we compare the workers' input in such questionnaires, and specifically on GMSI, to the music perceptual skills they might possess, which we measure through an audio-based, music perception skill-test.

[4] https://shiny.gold-msi.org/gmsiconfigurator/

[5] These values indicate the validity and reliability of the measurement tool when considering a set of questionnaire questions.

Subscale	Items	IRT Reliability	IRT Error	Reliability Alpha	Reliability Omega	Reliability Retest	CFI	TLI
Active Engagement	7	0.88	0.34	0.87	0.87	0.92	0.98	0.97
Perceptual Abilities	7	0.89	0.33	0.85	0.85	0.80	0.97	0.96
Musical Training	7	0.91	0.29	0.90	0.91	0.96	0.89	0.83
Singing Abilities	7	0.89	0.34	0.87	0.87	0.92	0.89	0.84
Emotions	6	0.83	0.41	0.79	0.79	0.85	0.95	0.92
General Music Sophistication	18	0.94	0.25	0.92	0.93	0.96	0.79	0.76

Reliability Alpha = Cronbach's Alpha; Reliability Omega = MacDonald's Omega; CFI = Comparative Fit Index; TLI = Tucker-Lewis Index.

Table 1: GMSI psychometric values

Measuring Music Perception Skills of Workers. The music perception skill test is based on the well-establish *Profile of Music Perception Skills* (PROMS) test (Law and Zentner 2012). Its original version is quite extensive, and its completion can take more than an hour, as it covers several music cognition aspects like Loudness, Standard rhythm, Rhythm-to-melody, Timbre, Pitch and more. Considering the possibly low familiarity of crowd workers with these tasks, and its inherent difficulty, we opted for a shorter version, the *Mini-PROMS* (Zentner and Strauss 2017), which has also been adopted and validate in the context of online, uncontrolled studies.

Mini-PROMS is a much shorter battery of tests (15 minutes completion time), which still covers the "Sequential" and "Sensory" subtests. It can measure a person's music perception skills, by testing their capability to indicate differences on the following musical features:

Melody: A sequence of notes, with varying density and atonality

Accent: The emphasis of certain notes in a rhythmic pattern

Tuning: The certain frequency of notes, when played in a chord

Tempo: The speed of a rhythmic pattern

The musical aspects selected in this test, are argued to well represent the overall music perception skills of a person, only in a more concise way. This version, retains test–retest reliability and internal consistency values close to the original PROMS test (Law and Zentner 2012), validating it for our research purposes. Note that, will reduced in size, these four skills are required to enable a broad range of music-related research, such as beat tracking, tonal description, performance assessment and more.

For each of the 4 musical aspects, workers receive a brief explanation and an example case to familiarise the user with the test. Each challenge after the introduction, presents a reference audio sample twice and a comparison sample once. The two audio samples can differ based on the musical aspect tested and the worker is asked if the samples are indeed same or differ. The authors of PROMS have put particular effort on distinguishing the musical aspects from each other, to make the skill evaluation as close as possible to the musical aspect tested. Finally, to minimise cognitive biases due to enculturation (Demorest et al. 2008), the audio samples have been created using less popular instrument sounds, such as harpsichord and "rim shots". Meanwhile, the structure of audio samples and the aspect separation allow for a more precise measurement of a person's perception skill.

The categories of "Melody" and "Accent" have 10 comparisons each, while "Tuning" and "Tempo" have 8. After the user has listened to the audio samples, they are asked to select between "Definitely Same", "Probably Same", "I don't know", "Probably Different" and "Definitely Different". The participant is then rewarded with 1 point for the high-confidence correct answer, while the low-confidence one rewards 0.5 point. The subscale scores are calculated through a sum of all items within the scale and divided by 2. The total score is an aggregated result of all subscale scores. During the test, the user is fully aware of the subscale they are tested for, but the name of "Tempo" is presented as "Speed" (original creators' design choice).

Self-assessment on Music Perception Skills. Self-assessment can often misrepresent an individual's real abilities(Kruger and Dunning 1999). For that reason, we employed a survey to study this effect its manifestation with music-related skills. After Mini-PROMS test, the worker has to input how many of the comparisons per subscale they believe they correctly completed - this information is not known to them after executing the Mini-PROMS test. Therefore, they are presented with 4 questions, where they have to indicate between 0 and the total number of tests per subscale (10 for "Melody"/"Accent" and 8 for "Tuning"/"Tempo"). Finally, the results of this survey, are compared to the score of workers on the "Perceptual Abilities" subscale of GMSI, which also relies on self-assessment. We expect workers to re-evaluate their own skills, once exposed to the perception skill test.

Post-task Survey. As a final step of the task, the worker is presented with three post-task surveys: 1) an survey on the audio equipment and the noise levels around them, 2) a survey on the cognitive load they perceived and 3) an open-ended feedback form.

The audio equipment survey consisted of four main questions, to retrieve the type of equipment, its condition and the levels of noise around them during the audio tests. Insights on these can help us understand the to what extent the equipment/noise conditions affected Mini-Proms test, which is audio-based. More specifically, we asked the following questions:

1. What audio equipment were you using during the music

skill test?

2. What was the condition of your audio equipment?
3. Does your audio equipment have any impairment?
4. How noisy was the environment around you?

The options regarding the audio equipment were: "Headphones", "Earphones", "Laptop Speakers" and "Dedicated Speakers". For the condition questions (2) and (3), we used the unipolar discrete five-grade scales introduced in (Recommendation 2003), to subjectively assess the sound quality of the participants' equipment. Finally, for question (4) on noise levels, we used the loudness subjective rating scale, introduced in (Beach, Williams, and Gilliver 2012).

In the second part of post-task survey, the workers had to indicate their cognitive task load, through the NASA's Task Load IndeX (NASA-TLX) survey[6]. The survey contains six dimensions — Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. Workers use a slider (ranging from 0 to 20, and later scaled to 0 to 100) to report their feelings for each of the six dimensions. A low TLX score represents the music skill test is not mentally, physically, and temporally demanding, and it also indicates less effort, and less frustration perceived by the worker, while completing the entire study.

Finally, we introduced an free-form textual feedback page, where users were encouraged to leave any comments, remarks, or suggestions for our study.

Worker Interface

The worker interfaces of our study is using VueJS[7], a JavaScript framework. The first page of our study, contained general instructions for the study alongside estimated completion times for each part of it. Each page thereafter, contained an interface for each of the steps in our study, as seen in Figure 1.

To assist navigation through the GMSI questionnaire, we implemented the questionnaire interface to show one question at a time. We added a small drifting animation to show the next question, when they select their answer in the previous one. We also added a "back" button, in case they wanted to return to a previous question and alter their answer. They could track their progress through the questionnaire from an indication of the number of the question and the total number of questions (see Figure 2).

While we retrieved the questions for GMSI and implemented them in our study's codebase, for PROMS we wanted to use the exact conditions and audio-samples as in (Zentner and Strauss 2017). To replicate their test faithfully, the creators of PROMS (Law and Zentner 2012) kindly gave us access to their Mini-PROMS interfaces (example interface in Figure 3). Mini-PROMS is implemented on LimeSurvey[8] and users were redirected to it after the completion of GMSI.

After the GMSI questionnaire, workers were introduced to the page seen in Figure 4. There, they had to copy their

[6]https://humansystems.arc.nasa.gov/groups/tlx/

[7]https://vuejs.org

[8]https://www.limesurvey.org

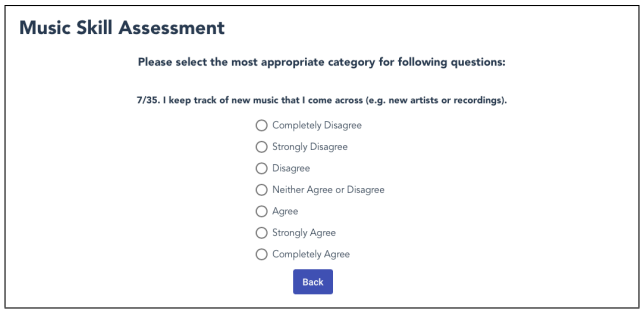

Figure 2: Interface of GMSI questionnaire.

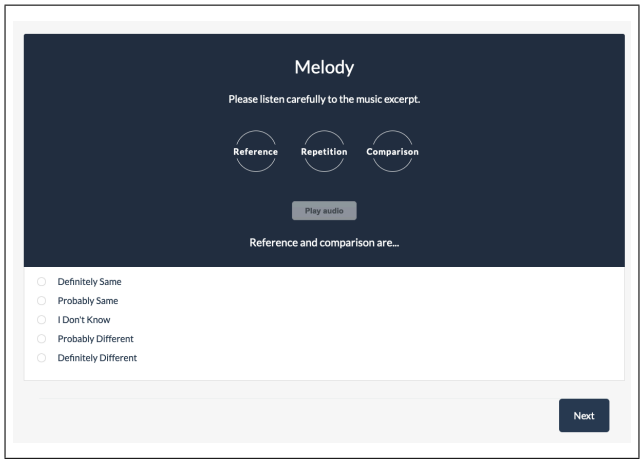

Figure 3: Interface of Mini-PROMS: "Melody" test.

Participant ID (retrieved programmatically from Prolific) and use it in the Mini-PROMS interface later, so we could link their test performance (stored in LimeSurvey), with their entries in our database. At the end of Mini-PROMS, the users were redirected back to our study through a provided URL.

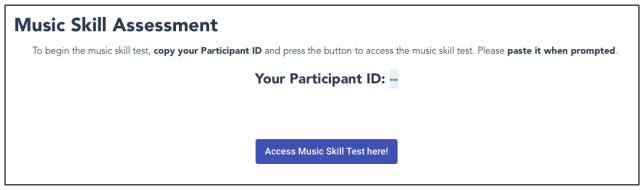

Figure 4: Redirecting page

In the final stage of our study, the participants were greeted and provided a "completion code", which they could submit on Prolific to complete the task.

Participants, Quality Control, and Rewards

On Prolific, we recruited 100 crowd workers to complete our study. We applied a participant selection rule for "Language Fluency": English, as all of our interfaces were implemented in English. Only crowd workers whose overall approval rates were higher than 90% could preview and perform our study.

To assess the quality of the user input, we included attention check questions on the GMSI and NASA-TLX interfaces of the study. More specifically, we included three attention check questions in GMSI, asking the participants to select a specific item in the same seven-point Likert scale. In the NASA-TLX survey, we included a question asking the users to select a specific value out of the 21 available in the scale of the survey. Of the 100 workers recruited from Prolific, 8 of them failed at least one attention check question(s); 5 of them provided invalid/none inputs. After excluding these 13 invalid submissions, we have 87 valid submissions from 87 unique workers.

We set the reward on Prolific for completing our study to 3.75 GBP (5.2 USD). Upon the completion of our study on Prolific, workers immediately received the reward. The average execution time was 32.5 minutes, resulting in the hourly wage of 7.5 GBP (10.3 USD), rated as a "good" pay by the platform.

Results

Worker Demographics

Table 2 summarises workers' demographic information. Of the 87 crowd workers who provided valid submissions, 36 were female (41.38%), while 51 were male (58.62%). Age of participants ranged between 18 and 58 and the majority of them were younger than 35 (87.36%). The majority of the workers (51%) were reported to be unemployed, while from those employed, 73.17% had a full-time job. Most workers had enrolled for or acquired a degree (78.16%), with 51.47% of them pointing to Bachelor's degree. In total, we employed workers from 15 countries, with most workers (77%) currently residing in Portugal (25), United Kingdom (16), Poland (13) and South Africa (13).

	Variables	**Statistics**
Gender, n (%)	Female	36 (41.38%)
	Male	51 (58.62%)
Age (years)	Range	18-58
	Mean (SD)	25 (7.8)
Occupation	Full-time	30
	Part-time	11
	Unemployed	44
	Voluntary Work	2
Education	Associate degree	3
	Bachelor's degree	35
	Doctorate degree	1
	High school/HED	16
	Master's degree	12
	Professional degree	1
	Some college, no diploma	13
	Some high school, no diploma	2
	Technical/trade/vocational training	4

Table 2: Participant demographics

Results on Worker Music Sophistication

Table 3 and Figure 5 summarise the results of the GMSI questionnaire on our workers. We contrast our results to results of the original GMSI study (Müllensiefen et al. 2014), which covered a large population sample of participants $n = 147,663$ that voluntary completed the questionnaire, on BBC's *How Musical Are You?* online test. Participants were mainly UK residents (66.9%) and, in general, from English-speaking countries (USA: 14.2%, Canada: 2.3%, Australia: 1.1%), with 15.9% having non-white background. The sample contained a large spread on education and occupation demographics, where only 1.8% claimed working in the music domain. To some extent, this study is considered representative for the general population in the UK (but is biased towards higher musicality due to the voluntary nature of that study). As such, we can assume a certain disposition and affinity to music from GMSI's population sample, compared to ours where the incentives where monetary.

In our study, the observed General Music Sophistication ($\mu = 69.76$) positions our workers pool at the bottom 28-29% of the general population distribution found in the GMSI study. We observe a similar effect also with the individual subscales with the exception of "Emotions", for which our workers fare a bit higher (bottom 32-38%).

	Range	Mean	Standard Deviation
Active Engagement	19-45	30.91	5.45
Perceptual Abilities	16-45	33.62	6.65
Musical Training	7-45	18.52	9.61
Singing Abilities	9-41	27.41	6.03
Emotions	18-42	33.24	4.28
General Music Sophistication	40-101	69.76	14.20

Table 3: GMSI Mean, Standard Deviation and Range

The result indicates that the self-reported music sophistication of crowd workers is strongly below that of the general population. Most workers had received relatively little formal training in their lifetime (with outliers of highly educated individuals, as seen in Figure 5). This finding is important for the rest of the analysis, as it indicates *low formal expertise* with music among the crowd workers.

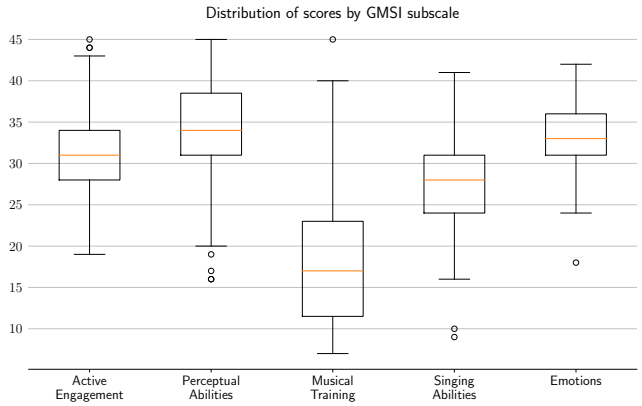

Figure 5: Distribution of GMSI subscale scores.

Most workers indicate relatively high perceptual abilities ($\mu = 33.62$, $max = 45$). Here, it is interesting that pre-

vious studies (Baharloo et al. 2000) estimate that less than 1% (or 5 people) per 11,000 possess "Absolute Pitch". In our sample though, 9 workers indicated having this characteristic, little more than the 10% of our sample. This could indicate a possible confusion between quasi-absolute pitch which is related to the familiarity of a person with an instrument's tuning and timber (Reymore and Hansen 2020), or with relative pitch. Relative pitch is trainable through practice and useful to professional musicians, as they can detect changes in pitch through the relations of tones (5 out of 9 workers who indicated "Absolute Pitch" had scored higher than 30 out of 49 in the "Musical Training" category scale, indicating adequate formal musical training).

Table 4 presents the correlations between GMSI subscales. As the scores of each GMSI subscale follow a normal distribution (Shapiro-Wilk test), we applied Pearson's R test to calculate correlation coefficients. We observe that Perceptual Abilities shows positive correlations with most other subscales ($p < 0.05$), especially with Music Training ($R = 0.442$), Emotions ($R = 0.380$), and Singing Abilities ($R = 0.463$). This finding suggests that the listening skill plays the most important role in crowd workers' music sophistication. We also find significant correlations between Active Engagement and Emotions ($R = 0.401$), and between Singing Abilities and Musical Training ($R = 0.465$). The original GMSI study has shown that different subscales are strongly correlated ($R > 0.486$). The difference we observe could be partly explained by the generally lower musical sophistication scores of the crowd workers in our pool.

	Active Engagement	Perceptual Abilities	Musical Training	Emotions	Singing Abilities
Active Engagement	1.000				
Perceptual Abilities	0.262*	1.000			
Musical Training	0.224*	0.442*	1.000		
Emotions	0.401*	0.380*	0.178	1.000	
Singing Abilities	0.142	0.463*	0.465*	0.125	1.000

Statistical significance ($p < 0.05$) is marked using an asterisk (*).

Table 4: Intercorrelations (Pearson's R) of subscales of GMSI scores.

Results on Objective Music Perception Skills

This section discusses the results of the Mini-PROMS active perception skill test, contributing to the answer of **RQ1**. Mini-PROMS categorizes perception skills as "Basic" if the total obtained score is lower than 18, "Good" if between 18 and 22.5, "Excellent" for values between 23 and 27.5, and "Outstanding" for values over 28 (Zentner and Strauss 2017). The original Mini-PROMS study covered a total $n = 150$ sample of participants, all recruited from the university of Innsbruck, via email. Most of the participants were students with at least one degree ($n = 134$), aged 27 on average.

We observed (see Table 5) an average of "Good" music perception skills for our workers ($\mu = 19.53$, avg. accuracy 54.25%). 48 out of 87 (55.17%) produced reasonably high accuracy in music skill tests (belonging to "Good" and better

	Range	Mean	Standard Deviation
Melody	1.5-9	4.98	1.59
Tuning	1-7.5	4.22	1.62
Accent	0-9.5	5.19	1.84
Tempo	1-8	5.14	1.59
Mini-PROMS Total	6-30	19.53	4.98

Table 5: Mini-PROMS Mean, Standard Deviation and Range

categories according to Mini-PROMS results). These figures are lower compared to the results of the original study (Zentner and Strauss 2017) ($\mu = 24.56$, 68.2% avg. accuracy), a fact that we account to the greater representation of *non-musician* in our workers pool (67.82%), compared to the participants of the original Mini-PROMS study (where only 38.67% identified as non-musicians). However, considering the low formal training amongst the surveyed workers, we consider this result an indication of the existence of useful and somewhat abundant auditory music perception skills among untrained workers. Especially, in the top 10% of workers, ranked according to their total Mini-PROMS values, several achieved quite high accuracy, between 73.6% and 83.3%, which would indicate perception skills between "Excellent" and "Outstanding" in Mini-PROMS's scale. In the following section we will analyse in greater detail the relationship between the measured music sophistication and the perception skills.

A similar trend towards lower performance compared to the original Mini-PROMS study can be observed across the other musical aspects: workers correctly identified melody differences with 49.77% avg. accuracy (original study: 64.3%), tuning differences with 52.73% avg. accuracy (original: 68%), accent difference with 51.95% avg. accuracy (original study: 61.5%), and tempo differences with 64.3% avg. accuracy (original study: 81.25%).

The result of the music skill tests is in-line with the result of self-reported music sophistication from GMSI, suggesting that when compared to the populations covered by previous studies, crowd workers generally possess less music perception skills. To deepen the analysis, we calculated the intercorrelation of Mini-PROMS subscales, and made comparison with the original study (Zentner and Strauss 2017). Since the Mini-PROMS scores across all the subscales follow normal distributions (Shapiro-Wilk tests(Hanusz, Tarasinska, and Zielinski 2016)), we carried out Pearson's R tests to get the correlation coefficients and corresponding p-values. We find statistical significance on all the intercorrelations. Especially, we find that workers' music skills related to melody are positively correlated with their accent- and tempo-related skills ($R = 0.551$ and $R = 0.514$ respectively), while accent and tempo also shows a moderate correlation ($R = 0.468$). In comparison with the original study, we do not observe large differences in the R values, while we did with the GMSI results. The results of the intercorrelation analysis suggests that worker melody, accent, and tempo skills are related with each other in our population too. This is a positive result, that suggests 1) the

applicability of this testing tool also on this population, and 2) the possibility of developing more compact tests for music perception skills, for workers' screening or task assignment purposes.

	Melody	Tuning	Accent	Tempo
Melody	1.000			
Tuning	0.363*	1.000		
Accent	**0.551***	0.336*	1.000	
Tempo	**0.514***	0.245*	**0.468***	1.000

Statistical significance ($p < 0.05$) is marked using an asterisk (*).

Table 6: Intercorrelations (Pearson's R) of subcategories of Mini-PROMS scores.

When focusing on the top 10% of workers, we observed an accuracy on "Melody" between 75% and 90% , while the top 5% scored higher than 85%. A person with "Absolute Pitch" would be expected to achieve high accuracy on this test. Only one person in the top 10% had indicated "Absolute Pitch", but their accuracy was one of the lowest in the group (75%). This could indicate that the person is more likely to not possess such a characteristic. For the subcategory of "Tuning", the top 10% achieved accuracy between 81.25% and 93.75%, while the top 5% scored higher than 87.5%. On "Accent", the top 10% reached accuracy between 80% and 95%. Finally, on the subcategory of "Tempo" we measured accuracy of 87.5% and 100% in the top 10%, while the top 5% achieved perfect score of 100%.

These results suggest the presence of a substantial fraction of workers possessing higher music perception skills than expected from their training, although differently distributed. For example, workers who perceived well changes in "Melody", didn't perform equally well on the other categories. This could indicate that music perception skills do not necessarily "carry over" from one music feature to the other; other workers will be good in perceiving changes in tempo, while others on tuning. This encourages the use of the appropriate set of tests, to identify potentially high performing annotators. Thus, if we take as example beat tracking annotation tasks, it would be more beneficial to focus on testing the rhythm-related perception skills, as the other categories have lower chance to capture the appropriate workers for the task.

Comparison between GMSI and Mini-PROMS

To answer **RQ2**, we compare the results between GMSI and Mini-PROMS by running Pearson's R correlation between the subcategories and total scores of GMSI and Mini-PROMS respectively (see Table 7).

We observe moderate positive correlations between Accent-related music skills with worker self-reported music sophistication in terms of Perceptual Abilities ($R = 0.405$), Musical Training ($R = 0.410$), and the General Music Sophistication score ($R = 0.420$). To a lesser degree, we see Melody-related skills to correlate with the same exact GMSI subscales, with equally high statistical significance.

	Melody	Tuning	Accent	Tempo	Mini-PROMS
Active Engagement	0.003	-0.088	-0.011	-0.067	-0.053
Perceptual Abilities	0.337*	0.026	**0.405***	0.311*	0.365*
Musical Training	0.323*	0.316*	**0.410***	0.248*	**0.437***
Emotions	0.086	-0.010	0.098	0.052	0.077
Singing Abilities	0.014	-0.066	0.274*	0.064	0.104
Absolute Pitch	-0.007	-0.034	0.129	-0.066	0.013
General Music Sophistication	0.247*	0.098	**0.420***	0.180	0.324*

Statistical significance ($p < 0.05$) is marked using an asterisk (*).

Table 7: Correlations (Pearson's R) between the subcategories of GMSI scores and the subcategories of Mini-PROMS scores.

Both music perception skill categories belong to the "Sequential" types of music features, showing that Perceptual Abilities, Musical Training and the General Music Sophistication scores can potentially indicate the performance on those skills.

Furthermore, we find that Musical Training reported at GMSI shows significantly positive correlation with the overall Mini-PROMS performance, representing the worker objective music skill in general.

We can observe a weak positive correlation in the remaining categories, which indicates that in the studied population, GMSI and Mini-PROMS are only loosely related. The almost correlation close to zero between "Active Engagement" and "Emotions" to the Mini-PROMS categories replicate findings of a study targeting a larger sample (Müllensiefen et al. 2014) (not in a crowdsourcing setting), where GMSI was compared to melody memory and beat perception skills and they reported also low correlation values between them. These results indicates that the extent to which workers engage with music (e.g. budget/time allocated and activity on online music-related forums) is not a predictor for a worker's accuracy on audio-based music tasks. The same would hold true for questions on how emotionally deep can the workers connect to music. In other words, it is not possible to trust the enthusiasm or commitment towards music that workers report, at it wouldn't be a good indication of their actual music perception skills.

Finally, the low or in cases negative correlation of "Absolute Pitch" with the Mini-PROMS categories validates our scepticism towards workers who claimed to possess "Absolute Pitch" abilities.

Figure 6 shows a scatter plot that describes the distribution of crowd workers distributed in terms of their music skill accuracy (Mini-PROMS) and their music training experience (GMSI). With this analysis, we would like to understand if our sampled crowd workers pools "hides" 1) **Musical Sleepers**, who are not formally trained but can accurately perform music-related tasks, and 2) **Sleeping Musicians**, who have received years of musical training, but perform poorly in music-related tasks. Recall from Table 3 that most of our workers had not received formal training (see Figure 6). Considering the presence of a substantial num-

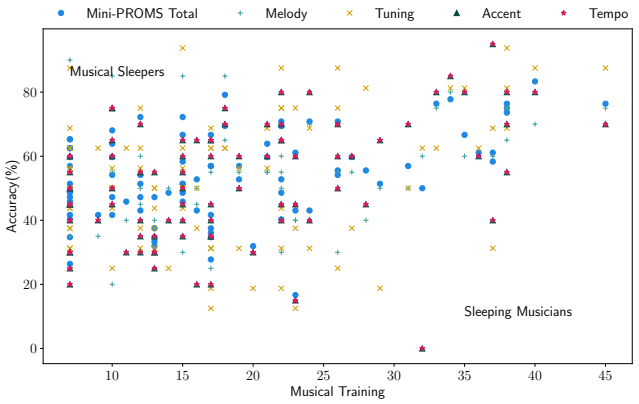

Figure 6: Musical Training (GMSI) and Performance on Mini-PROMS (acc%).

ber of highly performing workers (55.17% of the workers' pool), and the fact that very few workers received formal musical training (Figure 5), most of them can be qualified as proper music sleepers (Law and Zentner 2012). The presence of these workers is very encouraging, as it shows that it is possible to deploy advanced music analysis tasks on microtask platforms and and finding high-value contributors.

Self-assessed versus Measured Skill Levels

With this step, we study whether self-assessment can truly reflect one's music-related skills. The self-assessment accuracy and the actual Mini-PROMS accuracy are displayed together as a scatter plot in Figure 7. In this figure, blue points represent individual workers. Workers who located on the brown line perfectly self-assessed their performances while doing the Mini-PROMS test (meaning self-assessment accuracy equals to actual accuracy). Clearly, the majority of the crowd workers who participated our study consistently over-estimated their performances on music skill tests, irrespective of their actual music perception skills. We therefore observe a Dunning-Kruger effect, similarly to what has been found in other types of crowdsourcing tasks (Gadiraju et al. 2017a).

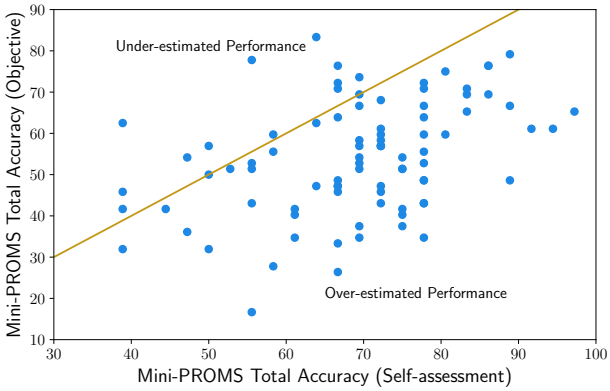

Figure 7: Distribution of Mini-PROMS total scores versus Self-assessed scores.

Post-task Survey: Equipment and Cognitive Workload

The majority of the workers reported that, during the test, they used headphones (52.87%) (which is very good for musical tasks), earphones (29.54%) and laptop speakers (16.09%) (which are not optimal). All workers reported the quality of their equipment as "Fair" or better quality (55.17% selected "Excellent" and 34.48% "Good"). 96.55% argued that their equipment either does not have any impairment (72.41%) or that the impairment is not annoying (24.13%). Finally, the majority of workers (58.62%) reported near silence conditions, while 31.03% of them reported normal, non-distracting levels of noise. While these conditions are not comparable to lab setups, we consider them to be sufficiently good to accommodate the requirements of our study.

We study workers' cognitive workload to understand if worker performance in music-related tasks is related to the perceived cognitive workload. We performed Pearson's R tests to calculate correlation coefficients between dimensions of cognitive worklaod (measured by NASA-TLX) and subscales of Mini-PROMS. Results are reported in Table 8 We observe negative correlations across all the dimensions and subscales, which suggests that workers who perform well in the music-related tasks also tend to perceive less workload, meaning they could feel less demanding (mentally, physically, or temporally), more successful, less difficult, and less frustrated. In traditional crowd tasks, workers need to pay more attention and put more effort to achieve high-quality performance. This result indicates that music-related tasks, high accuracy could be more "spontaneously" achieved, with less cognitive workload.

	Melody	Tuning	Accent	Tempo	Mini-PROMS
Mental Demand	-0.093*	-0.202	-0.124	-0.054	-0.159
Physical Demand	-0.204	-0.377	-0.230	-0.106	-0.307
Temporal Demand	-0.190	-0.301	-0.095	-0.115	-0.231
Performance	-0.125	-0.071	-0.269	-0.073	-0.185
Effort	-0.019*	-0.042*	-0.253	-0.102	-0.146
Frustration	-0.146	-0.169	-0.288	-0.215	-0.277

Statistical significance ($p < 0.05$) is marked using an asterisk (*).

Table 8: Correlations between worker cognitive workload and Mini-PROMS music skill test scores.

Discussion

In this study, we extensively measure the musical sophistication and music perception skills of crowd workers. We show that the self-reported music sophistication of crowd workers is below that of the general population and that formally-trained workers are rare. Nevertheless, we found surprisingly refined and diverse music perception skills amongst the top performers, which cannot accurately be predicted by questions regarding their engagement with music as a hobby. Studying the distribution of workers, we find evidence that supports the existence of workers with high accuracy and little to no formal training on crowdsourcing plat-

forms, namely "musical sleepers", indicating the prospect of high-quality annotations by non-experts on these platforms. However, this opens the challenge of how we can identify these "musical sleepers" reliably during worker selection.

Implications for Design

Self-reported Musical Sophistication. The musical sophistication assessments (GMSI) is a useful tool to evaluate workers' capability in completing music-related tasks. It is however a lengthy questionnaire, which could result in extra cost and worse worker engagement. Reducing the number of question is possible, but with implication in terms of test reliability. For instance, the subscale of Musical Training is positively correlated to their actual music perception skills (and the correlation coefficient is higher than the general GMSI). As music perception skills are of primary relevance when executing music-related tasks, we suggest that in future task design, requesters could consider using the subscale of musical training which only contains 7 items. This could be complemented with novel methods to effectively and precisely predict worker performance to further facilitate task scheduling and assignment.

Music Perception Skill Assessment. The Mini-PROMS tool appears to be an effective mean to evaluate worker quality in terms of music skills. Yet, it suffers from the same overhead issues of GMSI. In this case, we suggest to use PROMS or Mini-PROMS as a qualification test, possibly featured by crowdsourcing platforms. Workers could use this test to get the corresponding qualification, to obtain the opportunities to access more tasks, and earn more rewards.

Music Annotation and Analysis Tasks. The results of this study indicate that knowledge- and skill-intensive musical tasks could be deployed on microtaks crowdsourcing platforms, with good expectations in terms of availability of skilled workers. However, performance on different skills (Melody, Tuning, Accent, and Tempo) appears to be unevenly distributed. We therefore recommend to analyse the capabilities of the selected crowd and tailor the design of advanced music annotation and analysis tasks to precise music perception skills.

Limitations and Future Work

A main limitation of our study is concerned with the size and diversity of the tested population. A larger and/or more diverse participation pool could potentially aid the generalisability of our findings and lead to more fine-grained insights. Even though our results are based on a population of crowd workers that have received less formal musical training than the average population used in similar studies (Müllensiefen et al. 2014), the use of standardised and validate tests, lend confidence to the reliability of our findings.

Another potential confounding factor in our study, is the motivation for participation. We attracted crowd workers using monetary rewards, while in other studies people voluntarily performed their test (e.g. BBC's main Science webpage) (Müllensiefen et al. 2014). Such a difference could also explain the differences in observed distributions (musical training and perception skills). However, monetary incentives are a feature of crowdsourcing markets, which makes them appealing in terms of work capacity and likelihood of speedy completion. In that respect, our findings are very encouraging, as they show the availability of both musically educated and/or naturally skilled workers that could take on musically complex tasks.

As demonstrated in our results, workers who perform well in a certain perception category (e.g. "Melody") do not perform equally well in another (e.g. "Tempo"). In future studies, we encourage the use of perception tests, adjusted and adapted for the specific music task at hand by using the appropriate categories, to accurately select potentially highly performing workers.

Fatigue and distraction could have played a role in shaping the observed worker accuracy of the perception results. The relatively limited number of excluded workers (13%) though, and the limited cognitive workload experienced (on average) by workers, gives us an indication against this risk.

In this study, we utilized standardized tools to capture domain-specific characteristics of the workers of a specific platform. Comparing results from their self-reported "connection" to the domain, with those from actively testing their skills, can paint a clear picture of the workers' demographics on a specific domain. While this work is specific to the music domain, we believe that similar workflows can be utilized to study the characteristics of workers on other domains. This holds especially true, as crowdsourcing platforms have diverse user-bases and direct comparisons cannot safely be drawn to studies with highly controlled population samples.

Conclusion

In this paper, we have presented a study exploring the prevalence and distribution of music perception skills of the general crowd in an open crowdsourcing marketplace. We measured and compared self-reported musical sophistication and active music perception skills of crowd workers by leveraging the established GMSI questionnaire and Mini-PROMS audio-based test, respectively. Our analysis shows that self-reported musical sophistication of crowd workers is generally below the general population and the majority of them have not received any form of formal training. Despite that, we observed a substantial number of workers (55.17%) who achieved a reasonably high accuracy in music perception skill tests, alongside a substantial presence of *musical sleepers*. Moreover, our analysis shows worker accessibility to adequate equipment. Together, these findings indicate the possibility of further increasing the adoption of crowdsourcing as a viable means to perform complex music-related tasks. Future work will focus on conducting experiments with a larger and more diverse pool of workers (e.g. drawn from platforms like Amazon Mechanical Turk and Toloka), to gain further insights and improve the generalisability of our findings.

References

Baharloo, S.; Johnston, P. A.; Service, S. K.; Gitschier, J.; and Freimer, N. B. 1998. Absolute pitch: an approach for

identification of genetic and nongenetic components. *The American Journal of Human Genetics* 62(2): 224–231.

Baharloo, S.; Service, S. K.; Risch, N.; Gitschier, J.; and Freimer, N. B. 2000. Familial aggregation of absolute pitch. *The American Journal of Human Genetics* 67(3): 755–758.

Beach, E. F.; Williams, W.; and Gilliver, M. 2012. The objective-subjective assessment of noise: Young adults can estimate loudness of events and lifestyle noise. *International journal of audiology* 51(6): 444–449.

Burkhard, A.; Elmer, S.; and Jäncke, L. 2019. Early tone categorization in absolute pitch musicians is subserved by the right-sided perisylvian brain. *Scientific reports* 9(1): 1–14.

Degrave, P.; and Dedonder, J. 2019. A French translation of the Goldsmiths Musical Sophistication Index, an instrument to assess self-reported musical skills, abilities and behaviours. *Journal of New Music Research* 48(2): 138–144.

Demorest, S. M.; Morrison, S. J.; Jungbluth, D.; and Beken, M. N. 2008. Lost in translation: An enculturation effect in music memory performance. *Music Perception* 25(3): 213–223.

Gadiraju, U.; Fetahu, B.; Kawase, R.; Siehndel, P.; and Dietze, S. 2017a. Using worker self-assessments for competence-based pre-selection in crowdsourcing microtasks. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24(4): 1–26.

Gadiraju, U.; Möller, S.; Nöllenburg, M.; Saupe, D.; Egger-Lampl, S.; Archambault, D.; and Fisher, B. 2017b. Crowdsourcing versus the laboratory: towards human-centered experiments using the crowd. In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*, 6–26. Springer.

Gingras, B.; Honing, H.; Peretz, I.; Trainor, L. J.; and Fisher, S. E. 2015. Defining the biological bases of individual differences in musicality. *Philosophical Transactions of the Royal Society B: Biological Sciences* 370(1664): 20140092.

Hannon, E. E.; and Trainor, L. J. 2007. Music acquisition: effects of enculturation and formal training on development. *Trends in cognitive sciences* 11(11): 466–472.

Hanusz, Z.; Tarasinska, J.; and Zielinski, W. 2016. Shapiro-Wilk test with known mean. *REVSTAT-Statistical Journal* 14(1): 89–100.

Hulin, C.; Netemeyer, R.; and Cudeck, R. 2001. Can a reliability coefficient be too high? *Journal of Consumer Psychology* 55–58.

Hyde, K. L.; and Peretz, I. 2004. Brains that are out of tune but in time. *Psychological science* 15(5): 356–360.

Joshi, A.; Kale, S.; Chandel, S.; and Pal, D. K. 2015. Likert scale: Explored and explained. *British Journal of Applied Science & Technology* 7(4): 396.

Koelsch, S.; Schulze, K.; Sammler, D.; Fritz, T.; Müller, K.; and Gruber, O. 2009. Functional architecture of verbal and tonal working memory: an FMRI study. *Human brain mapping* 30(3): 859–873.

Kruger, J.; and Dunning, D. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology* 77(6): 1121.

Law, E. L.; Von Ahn, L.; Dannenberg, R. B.; and Crawford, M. 2007. TagATune: A Game for Music and Sound Annotation. In *ISMIR*, volume 3, 2.

Law, L. N.; and Zentner, M. 2012. Assessing musical abilities objectively: Construction and validation of the Profile of Music Perception Skills. *PloS one* 7(12): e52508.

Lee, J. H. 2010. Crowdsourcing Music Similarity Judgments using Mechanical Turk. In *ISMIR*, 183–188.

Lee, J. H.; Hill, T.; and Work, L. 2012. What Does Music Mood Mean for Real Users? In *Proceedings of the 2012 IConference*, iConference '12, 112–119. New York, NY, USA: Association for Computing Machinery. ISBN 9781450307826. doi:10.1145/2132176.2132191. URL https://doi.org/10.1145/2132176.2132191.

Lee, J. H.; and Hu, X. 2012. Generating ground truth for music mood classification using mechanical turk. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, 129–138.

Liberman, A. M.; and Mattingly, I. G. 1985. The motor theory of speech perception revised. *Cognition* 21(1): 1–36.

Lima, C. F.; Correia, A. I.; Müllensiefen, D.; and Castro, S. L. 2020. Goldsmiths Musical Sophistication Index (Gold-MSI): Portuguese version and associations with socio-demographic factors, personality and music preferences. *Psychology of Music* 48(3): 376–388.

Lin, H.-R.; Kopiez, R.; Müllensiefen, D.; and Wolf, A. 2021. The Chinese version of the Gold-MSI: Adaptation and validation of an inventory for the measurement of musical sophistication in a Taiwanese sample. *Musicae Scientiae* 25(2): 226–251.

Mandel, M. I.; Eck, D.; and Bengio, Y. 2010. Learning tags that vary within a song. *ISMIR, Utrecht, Netherlands* .

Mankel, K.; and Bidelman, G. M. 2018. Inherent auditory skills rather than formal music training shape the neural encoding of speech. *Proceedings of the National Academy of Sciences* 115(51): 13129–13134.

Müllensiefen, D.; Gingras, B.; Musil, J.; and Stewart, L. 2014. The musicality of non-musicians: an index for assessing musical sophistication in the general population. *PloS one* 9(2): e89642.

Oh, J.; and Wang, G. 2012. Evaluating crowdsourcing through amazon mechanical turk as a technique for conducting music perception experiments. In *Proceedings of the 12th International Conference on Music Perception and Cognition*, 1–6.

Oosterman, J.; Yang, J.; Bozzon, A.; Aroyo, L.; and Houben, G.-J. 2015. On the impact of knowledge extraction and aggregation on crowdsourced annotation of visual artworks. *Computer Networks* 90: 133 – 149. ISSN 1389-1286. doi:https://doi.org/10.1016/j.comnet.2015.07.

008. URL http://www.sciencedirect.com/science/article/pii/S1389128615002315. Crowdsourcing.

Peretz, I.; and Hyde, K. L. 2003. What is specific to music processing? Insights from congenital amusia. *Trends in cognitive sciences* 7(8): 362–367.

Recommendation, I. 2003. General methods for the subjective assessment of sound quality. *ITU-R BS* 1284–1.

Reymore, L.; and Hansen, N. C. 2020. A theory of instrument-specific absolute pitch. *Frontiers in psychology* 11: 2801.

Samiotis, I. P.; Qiu, S.; Mauri, A.; Liem, C. C. S.; Lofi, C.; and Bozzon, A. 2020. Microtask crowdsourcing for music score Transcriptions: an experiment with error detection. In *Proceedings of the 21st International Society for Music Information Retrieval Conference*.

Schaal, N. K.; Bauer, A.-K. R.; and Müllensiefen, D. 2014. Der Gold-MSI: replikation und validierung eines fragebogeninstrumentes zur messung musikalischer erfahrenheit anhand einer deutschen stichprobe. *Musicae Scientiae* 18(4): 423–447.

Sorokin, A.; and Forsyth, D. 2008. Utility data annotation with amazon mechanical turk. In *2008 IEEE computer society conference on computer vision and pattern recognition workshops*, 1–8. IEEE.

Speck, J. A.; Schmidt, E. M.; Morton, B. G.; and Kim, Y. E. 2011. A Comparative Study of Collaborative vs. Traditional Musical Mood Annotation. In *ISMIR*, volume 104, 549–554. Citeseer.

Totterdell, P.; and Niven, K. 2014. *Workplace moods and emotions: A review of research*. Createspace Independent Publishing.

Ullén, F.; Mosing, 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: 87–93.

Urbano, J.; Morato, J.; Marrero, M.; and Martín, D. 2010. Crowdsourcing preference judgments for evaluation of music similarity tasks. In *ACM SIGIR workshop on crowdsourcing for search evaluation*, 9–16. ACM New York.

Werner, P. D.; Swope, A. J.; and Heide, F. J. 2006. The music experience questionnaire: Development and correlates. *The Journal of psychology* 140(4): 329–345.

Zentner, M.; and Strauss, H. 2017. Assessing musical ability quickly and objectively: development and validation of the Short-PROMS and the Mini-PROMS. *Annals of the New York Academy of Sciences* 1400(1): 33–45.

Zhuang, M.; and Gadiraju, U. 2019. In What Mood Are You Today? An Analysis of Crowd Workers' Mood, Performance and Engagement. In *Proceedings of the 10th ACM Conference on Web Science*, 373–382.