Journal of Research in Personality 75 (2018) 94–102

Contents lists available at ScienceDirect

Journal of Research in Personality

journal homepage: www.elsevier.com/locate/jrp

Registered Report

Computer-based music feature analysis mirrors human perception and can be used to measure individual music preference

Kai R. Fricke [a,*], David M. Greenberg [b,d], Peter J. Rentfrow [c], Philipp Yorck Herzberg [a]

[a] Department of Personality Psychology and Psychological Assessment, Helmut-Schmidt-University/University of the German Federal Armed Forces Hamburg, Hamburg, Germany
[b] The School of Performance, Anglia Ruskin University, Cambridge, United Kingdom
[c] Department of Psychology, School of the Biological Sciences, University of Cambridge, Cambridge, United Kingdom
[d] Autism Research Centre, Department of Psychiatry, University of Cambridge, Cambridge, United Kingdom

ARTICLE INFO

Article history:
Received 28 February 2018
Revised 6 June 2018
Accepted 8 June 2018

Keywords:
Music preference
Music features
Music taste
Music perception
Music information retrieval

ABSTRACT

This paper explores the measurement of individual music feature preference using human- and computer-rated music excerpts. In the first of two studies, we correlated human ratings of song excerpts with computer-extracted music features and found good accordance, as well as similar criterion validity with preference for musical styles (the MUSIC model, mean r = 0.88). In a second online study (N = 2118), using PCA and Procrustes analysis, we found that measured music preference showed the same established three-component structure from previous research (Arousal, Valence, Depth), regardless of whether the music pieces were rated by humans or the ESSENTIA music analysis software. Our results suggest that computer-extracted music features can be used to assess individual music preference.

1. Introduction

Research on musical genre and feature preference has established that humans show individual music preferences (e.g. Rentfrow, Goldberg, & Levitin, 2011), which in turn are related to various personality traits, such as Sensation Seeking (Litle & Zuckerman, 1986), the Big Five personality dimensions (Rentfrow & Gosling, 2003), and their facets (Zweigenhaft, 2008; Greenberg et al., 2016). Besides its social implications, such as acting as a conveyor for social affiliation (North & Hargreaves, 1999), music preference thus can also be construed as an idiosyncratic personality characteristic. The assessment of music preferences currently often relies on audio based assessment, which in turn relies on a selection of pre-rated music pieces. Software-based rating of music pieces can be used to broaden the selection of stimuli for music preference assessment. Concretely, such automated ratings could be used to measure music preference from actual, individually selected or even user-provided songs (e.g. own music collections, playlists, and alike), rather than from a pre-selected pool of music. However, research has not yet examined whether these computer-extracted features are empirically equivalent to human perception. Specifically, while we can assume that computer-extracted features partly reflect human perception, it is unclear if the results can also be used to infer idiosyncratic music preference. Therefore, in this paper we examine how the results from computational music feature extraction relate to human perception and how they can be utilized for the measurement of individual music preference.

In research, music preference is usually assessed in one of two ways: Participants can state their preference for musical genres (e.g. Rentfrow & Gosling, 2003; Bonneville-Roussy, Rentfrow, Xu, & Potter, 2013) or musical features (e.g. Fricke & Herzberg, 2017) in self-report, or participants can express their liking of certain selected music excerpts, which are used as a proxy measure to assess their music preference (e.g. Rentfrow et al., 2011; Langmeyer, Guglhör-Rudan, & Tarnai, 2012). Much of the past research on excerpt-based music preference assessment relied on a pool of neatly compiled music excerpts that have been rated by humans on various music features (e.g. Rentfrow et al., 2011; Rentfrow et al., 2012; Greenberg et al., 2016). These excerpts were selected from professionally produced, yet commercially unreleased music pieces (Rentfrow et al., 2011). Over all, 250 song excerpts have been introduced to this line of research, and they have been rated by humans on over 40 sonic and psychological music features. Separate participants would listen to the excerpts and express their preference, which would then be used to calculate their music preference profile for the music features.

* Corresponding author.
E-mail address: fricke@hsu-hh.de (K.R. Fricke).

https://doi.org/10.1016/j.jrp.2018.06.004
0092-6566/

The measurements were then used to examine relationships with cognitive styles (Greenberg, Baron-Cohen, Stillwell, Kosinski, & Rentfrow, 2015) and personality constructs (Greenberg et al., 2016).

These music excerpts are suitable for the fine assessment of music preference in overt assessment. But if we wanted to infer music preference from indirect data sources, such as actual listening behavior, we would need access to the music features of a large body of popular music. While today it is easy to get access to a large selection of music through online streaming services, gaining human ratings of thousands of songs on multiple features is a challenging and very time-consuming task if done manually. A way to overcome this limitation is to use automatic music feature extraction from music analysis software. This software can infer low-level music features directly from audio input and thus automatically annotate thousands of songs in relatively short time.

Low-level audio features derived from analysis software can be subjected to machine learning algorithms to classify songs on more high-level features, such as moods (e.g. *Danceable*, *Aggressive*) or genres (e.g. *Rock*, *Pop*, *Jazz*). These algorithms generalize learned features from ground truth datasets (i.e. manually annotated music pieces) to new audio inputs. The ground truth data is usually annotated by humans; still, it is sensible to confirm the validity of the computed feature output on a different dataset using human ratings. Also, audio classifiers usually use relatively broad categories, such as *Happy*, or *Sad*. Comparing the machine ratings with human ratings on more detailed music features can help us understand the facets of such categories and help in the interpretation of the music analysis results and music preference assessment.

Music preference research shows a robust five-factor (for musical genres; see Rentfrow et al., 2011) and three-factor (for musical features; see Greenberg et al., 2016; Fricke & Herzberg, 2017) structure. If computer-extracted music features proved to be a valid way to assess music feature preference, we would expect to find a similar factor structure here, as well.

In our first study we use rating data based on human perception of 150 songs collected by Rentfrow et al. (2011, 2012) to validate the feature extractions of the music analysis software ESSENTIA (Bogdanov et al., 2013). In our second study, we then examine and compare the component structures of music feature preference for human-rated and computer-extracted music features. Our study aims to show that computer-extracted music features can be used for the assessment of music preference, just as human-rated music features have shown in previous research. The overall goal of this paper is hence to provide a new tool for the assessment of individual music preference, which can be applied to any body of music, including individually selected or user-provided music pieces.

1.1. Music features and music preference

Music can be described and classified using various dimensions. Human-driven classification often revolves around categories, such as genres, moods, and sound characteristics. Machine-extracted features on the other hand are typically very technical, as they usually describe quantitative characteristics of the audio signal, such as amplitudes, energies, and spectral bands. These technical attributes are referred to as *low-level* features. Classifications derived from a combination of these low-level features often mimic human concepts and are referred to *high-level* features. We will first discuss human classification and rating of music attributes, and then proceed to illustrate machine-extracted low-level and high-level features.

Traditionally, music has been categorized into genres such as *Rock*, *Pop*, or *Jazz*. Psychological research suggests that preference for these genres can be modeled on the five factors of *Mellow*, *Unpretentious*, *Sophisticated*, *Intense*, and *Contemporary* (the *MUSIC*-Model; Rentfrow et al., 2011). In the past years, however, research switched focus from broader music *genre* preference (e.g. "I like Jazz") to music *feature* preference (e.g. "I like relaxing music") (Fricke & Herzberg, 2017). Preference for music features in excerpt-based assessment has been shown to load on three factors: *Arousal*, *Valence*, and *Depth* (AVD) (Greenberg et al., 2016). This factor structure has been replicated for self-reported music preference assessment (Fricke & Herzberg, 2017).

The relationship of preference for musical styles (i.e., the MUSIC model dimensions) and for music features (i.e., the AVD model) has been examined in research (e.g. Rentfrow et al., 2012; Fricke & Herzberg, 2017). Furthermore, the factor structure of the MUSIC model has been confirmed in various studies, including a large cohort study with over 250,000 participants (Bonneville-Roussy et al., 2013). The relationship of music feature preference with the MUSIC model is hence suitable to determine criterion validity of the computer-extracted features.

Regarding the relationship to personality, music preference showed robust relationships to the Big Five personality dimensions (Fricke & Herzberg, 2017; Langmeyer et al., 2012; Zweigenhaft, 2008; Rentfrow & Gosling, 2003). A recent meta analysis confirmed some of these findings, but noted that most relationships are rather small (Schäfer & Mehlhorn, 2017). In addition to their primary relationships with music preference, personality traits have been shown to moderate age trends in music preference (Bonneville-Roussy et al., 2013). Further, some studies showed that music preference correlated with the same biological indicators as the Big Five personality traits; for instance, a higher testosterone level correlated negatively with preference for sophisticated music in males (Doi, Basadonne, Venuti, & Shinohara, 2018). All these results suggest that music preference itself is closely related to personality. As such, the exploration of novel assessment methods for music preference can be used to enable research and real-world applications to infer personality characteristics from music preference data, and use these insights to tailor their tasks and services to each user. Additionally, personality research suggested that digitally derived behavior data, such as Facebook likes, predicted personality better than judgments of human peers, and sometimes even self-ratings (Youyou, Kosinski, & Stillwell, 2015). It is hence conceivable that computer-based methods might provide increased validity, or at least increased objectivity over human rating-based assessments.

1.2. Automatic music feature extraction

The subfield of computer sciences that deals with extracting information from music data is called *Music Information Retrieval* (MIR). MIR systems seek to analyze music files in terms of pitch, tempo, harmony, and timbre, as well as editorial, textual, and bibliographic facets (Downie, 2003). The last two decades saw various research studies developing algorithms to extract these facets from music files, as well as the creation of several software suites implementing these algorithms.

Among these software suites are open-source solutions such as *librosa* (McFee et al., 2015), *jMIR* (McKay, 2010), *ESSENTIA* (Bogdanov et al., 2013), and many others. We chose to use ESSENTIA for our analyses, because (a) it is actively developed and maintained, (b) it implemented state-of-the-art MIR algorithms, and (c) it provides learned models for high-level features based on a large count of research papers.

ESSENTIA analyzes the digital audio data from songs and extracts various low-level parameters, such as *beats per minute*, *spectral complexity* and *MFCC* (Bogdanov et al., 2013). These low-level parameters can be administered to machine learning algorithms, which map certain low-level audio profiles to high-level features. ESSENTIA provides pre-trained Support Vector Machines

(SVM) for the automatic extraction of high-level music features (Bogdanov et al., 2013). These features include sonic (e.g. *acoustic*, *electronic*, *tonal/atonal*, *instrumental*) and psychological features (e.g. *aggressive*, *happy*, *sad*, *relaxed*, and *party music*), as well as rhythm and genre classifiers (Bogdanov et al., 2013). We included all available high-level mood and sound classifiers in our analysis, as well as five broader mood cluster classifiers, as they were all derived from high-quality ground truth data and proved their accuracy in previous research (Bogdanov, 2013). We also included some low-level features in our analysis which were directly related to sound features. Specifically, these were *Average loudness*, *Dissonance*, *Dynamic complexity*, and *Speed (Beats-per-Minute; BPM)*. Lastly, we included the results from the *Rosamerica* genre classifier (Bogdanov, 2013) for supplementary analysis.

In the annotation of ESSENTIA's ground truth data, experts confirmed the correctness of the tags, which were then used to learn the models. By administering these models to a new set of music files that has been annotated by different raters, we thus confirm and strengthen the validity of the models. Additionally, the human ratings have a higher level of granularity, enabling us to examine which facets are covered by ESSENTIA's high-level SVMs. Lastly, by collecting preference data we examine if we can replicate the three-factor structure of music feature preference with computer-extracted music features.

1.3. Aims

This paper aims to answer the following questions: (a) How do computer-extracted music features relate to human perception? (Study 1), (b) Are computer-extracted music features valid? (Study 1), (c) Can computer-extracted music features be used to measure music preference? (Study 2), and (d) How robust is the component model of music feature preference measured by computer-extracted music features? (Study 2).

2. Study 1: Equivalence of human ratings and computationally extracted music features

In the first study, we sought to verify the validity and explore the facets of computationally extracted features by comparing them to human ratings of the same songs. Further, we examined the external validity by comparing the correlations of computer-extracted and human-rated music features with the MUSIC model.

2.1. Method

2.1.1. Music excerpts

We examined the same 15 s music excerpts as Rentfrow and colleagues (Rentfrow et al. (2011);Rentfrow et al. (2012)) used in previous studies. To ensure that each excerpt was rated within the same frame of reference, we only included those excerpts which have been used in studies examining all genres (i.e. not those which were used to examine the MUSIC model within the Jazz and Rock genres in Rentfrow et al., 2012). In total, we had access to human ratings on 150 excerpts.[1] All excerpts were from a mixture of 26 genres (see Rentfrow et al., 2011). Power analysis revealed that this sample size is suitable to detect correlations of at least $r \geqslant 0.25$ with $\alpha = 0.05$ and $\beta = 0.20$ (minimum sample size N = 123; calculated using formula from Hulley, Cummings, Browner, Grady, & Newman, 2013, p. 79).

[1] The excerpts are freely available at: http://daniellevitin.com/levitinlab/LabWebsite/expsupport/MUSIC/Rentfrow_JPSP_Index.html.

2.1.2. Human ratings

The excerpts were rated on 45 selected music features by judges with no former music training (see Rentfrow et al., 2012; Greenberg et al., 2016). Ratings on all songs were available for the music features *instrumental*, *fast*, *loud*, *acoustic*, *percussive*, *dense*, *distorted*, *aggressive*, *romantic*, *sad*, *complex*, *relaxing*, *intelligent*, *inspiring*, as well as various other features for a selection of the excerpts.

2.1.3. Computer-extracted features

Music features were extracted from the raw audio data using the ESSENTIA software library (Bogdanov et al., 2013). ESSENTIA provides an out-of-the-box extractor for low-level and high-level features. High-level feature output in ESSENTIA is given as the probability of a song to belong to the respective class. For instance, a value of 0.90 on the *happy* output indicates a 90% chance that the song belongs to the *happy* category, as opposed to a 10% chance to not belong to this category. Note that not fitting into the *happy* category not necessarily means the song fits into the *sad* category.

2.1.4. Analysis

To examine the validity of the extracted features from ESSENTIA, we calculated Pearson correlations with the human rating data. We selected those ESSENTIA classifiers whose descriptions directly matched human-rated features (e.g. we compared ESSENTIA's *Relaxed* classifier with the human-rated *Relaxing* feature). Also, we examined which human features showed the largest positive and negative correlations with the ESSENTIA classifiers. To address criterion validity, we examined the correlations of the extracted features with the MUSIC model and compared them with those reported by previous studies.

2.2. Results

Unless otherwise noted, all correlations reported in the results section are significant with at least $p < .01$.

2.2.1. Relationship of human ratings with mood and sound classifiers

Analyzing the matching pairs of features between ESSENTIA and the ratings from Rentfrow et al. (2011, 2012) revealed correlations between $r = 0.21$ (*Happy*; $p = .03$) and $r = 0.65$ (*Party music*, $p < .001$), as seen in Table 1. We examined correlations with other features exploratively and found that they mostly fit well regarding their contents. For instance, ESSENTIA's *Aggressive* classifier correlated positively with the human-rated *Auditory* features, such as *Fast* ($r = 0.44$), *Loud* ($r = 0.63$) and *Percussive* ($r = 0.52$). The largest positive and negative correlations with the ESSENTIA classifiers from the primary analysis are presented in Table 1. Correlation tables of all human ratings and the ESSENTIA mood, sound an genres classifiers are available in the supplementary material (Tables B1 and B2).

2.2.2. Correlations with the MUSIC factors

Rentfrow and colleagues (Rentfrow et al., 2012) provided factor loadings on the MUSIC model for music pieces and human-rated features. First, we correlated the extracted features from ESSENTIA of the provided song excerpts with the loadings on the MUSIC model. In this case we used the music classifiers that best fit the respective human-rated features (e.g. *Electronic* and *Electric*). Then, we compared these correlations with the feature loadings provided by Rentfrow et al. (2012) using column-vector correlations, as seen in Table 2.

The correlations between the ESSENTIA features and the MUSIC factors revealed similar relationships as the research from Rentfrow et al. (2012). In fact, when comparing the correlations, we found column-vector correlations between $r = 0.84$ and

Table 1
Correlations between matching features of ESSENTIA features and human ratings

					Largest negative correlation[a]		Largest positive correlation[a]	
ESSENTIA classifier	Human feature	r	n	p	Human feature	r	Human feature	r
Danceable	Danceable	0.32	102	.001	Thoughtful	−0.40	Heavy bass[b]	0.73
Sad	Sad	0.39	150	<.001	Wild	−0.68	Gentle	0.71
Happy	Happy	0.21	103	.03	Calming	−0.36	Danceable	0.35
Relaxed	Relaxing	0.59	150	<.001	Abrasive	−0.65	Calming	0.71
Party	Party music	0.65	102	<.001	Gentle	−0.64	Heavy bass[b]	0.69
Instrumental	Instrumental	0.62	150	<.001	Raspy voice[b]	−0.47	Sophisticated	0.61

Note. (a) All p <.001, and all n = 102, except for (b) n = 50.

Table 2
Correlations between the MUSIC factors and music features.

	Mellow		Unpretentious		Sophisticated		Intense		Contemporary	
	E	R	E	R	E	R	E	R	E	R
Aggressive	−0.57	−0.62	−0.17	−0.29	−0.50	−0.41	0.85	0.78	−0.21	−0.14
Danceable	−0.18	−0.37	−0.24	0.05	−0.40	−0.35	0.04	0.08	0.63	0.43
Electric	−0.27	−0.23	−0.42	−0.40	−0.20	−0.57	−0.02	0.38	0.77	0.52
Happy	−0.17	−0.04	0.29	0.18	−0.22	0.24	0.1	−0.34	−0.14	0.18
Party music	−0.45	−0.55	−0.32	−0.20	−0.60	−0.49	0.43	0.44	0.50	0.41
Relaxing	0.51	0.65	0.05	0.06	0.62	0.53	−0.79	−0.61	0.13	−0.05
Sad	0.60	0.35	0.17	0.23	0.46	0.17	−0.6	−0.21	−0.08	−0.26
Instrumental	0.27	0.20	−0.38	−0.47	0.51	0.28	−0.18	0.09	−0.02	0.01
CVC	0.95		0.87		0.84		0.85		0.87	

Note. CVC = Column vector correlations, E = features from ESSENTIA, R = human ratings from Rentfrow et al. (2012).

r = 0.94, indicating a high similarity of both feature collection methods in terms of criterion validity.

2.3. Discussion

Comparison of the automatically extracted music features using ESSENTIA and the human ratings revealed a good accordance of both rating methods. Correlations between matching pairs of features were mostly of medium magnitude, with the exception of the *Happy* classifier, which showed smaller accordance. The largest negative and positive correlations with the respective ESSENTIA classifiers were found to be sensible, e.g., a correlation of r = −0.68 between *Sad* and *Wild* indicated that music that ESSENTIA classified as *Sad* is perceived as not *Wild* by human raters. The results indicate that both methods measure similar dimensions, which is once more reinforced by the very high similarity between the human- and machine-rating method in terms of criterion validity. Specifically, the correlation pattern with the MUSIC model is so similar between both methods that we can assume that both human ratings and machine ratings describe the five MUSIC dimensions in the same way.

Although machine learned classifiers might not provide the same accuracy as human ratings, our results indicate that they are valid and can be used for broad music feature extraction. Automatic feature extractors thus provide a valid way for automatically annotating music and are suitable for large-scale music analysis.

3. Study 2: Factor structure of music feature preference

With the indication that music features can be appropriately retrieved using MIR software, the question arises if such features can be used to measure individual music preference. In the second study, we therefore collected music preference data and music feature ratings from a large online sample. This allowed us to examine and compare the component structure of music feature preference measured using either human-rated or machine-extracted feature ratings. Further, by using two separate lists of songs for different participants, we can examine the robustness of the component model for both methods.

In the remainder of this text we will refer to the previously obtained ratings by Rentfrow et al. (2012) as the *MUSIC-Model* data, and the newly collected ratings as the *Musical Universe* (*MU*) data.

3.1. Method

Music feature human-rating data was obtained from the *Musical Universe* project. The Musical Universe is a website (www.musicaluniverse.org) that collects data from participants in exchange for feedback on their results. The project was featured in various news outlets (e.g. CNN, IFLscience.com) and administered different supplemental questionnaires over its time of existence.

In our study, participants were provided with 10 music excerpts (two for each MUSIC factor) in two versions. The songs were selected from the excerpt pool (see Rentfrow et al., 2011) as described in Study 1. The songs were selected so that they matched the genre composition of the total song pool. The list of excerpts can be found in the supplementary material in Table B3.

After answering a set of demographic questions, subjects were presented with one excerpt at the time. They listened to the excerpt and were then asked to rate it on 29 features, as well as to state their preference for the excerpt. With 10 excerpts per version this resulted in 300 items per subject. For our analysis, we only considered subjects who answered at least 120 (40%) of the items. Also, we excluded participants older than 65 years or younger than 18 years to avoid impairment from hearing difficulties. This totaled to 2118 participants (1987 in version A, 131 in version B).

Twenty-five music features were taken from Greenberg et al. (2016). A further four items have been initially added to the questionnaire, but have not been included in the analysis as they were ambiguous and not derived from theory or analysis like the other items.

Participants had a mean age of 31.56 years (*SD* = 11.47). Sex was almost evenly distributed, with 880 (41.5%) males and 1017 (48.0%) (a further 13 identified as *transgender*, and another 8 as *other* while the remainder did not answer the question). The majority of the participants (1466, or 69.2%) were of white Caucasian ethnicity (4.5% Latino, 3.7% Mixed, 3.4% Chinese, 9.8% did not answer). Regarding their origin, most participants came from English-speaking countries, i.e. 656 participants (31.0%) resided in the USA, 281 (13.3%) in the UK, 152 (7.2%) in Canada, and 107 (5.1%) in Australia. Further 222 (10.5%) did not state their country of origin. The rest is distributed among 67 other countries.

On average, participants reported the importance of music in their lives as 5.97 (*SD* = 1.07) on a 7-item scale, corresponding to *very important*. About half participants stated they played a musical instrument (927, as opposed to 949 not playing an instrument and 242 who did not answer this question).

The minimum sample size to detect correlations of at least $r \geqslant 0.25$ with $\alpha = 0.05$ and $\beta = 0.20$ remains N = 123. Power analysis revealed that our much larger sample is able to identify correlations under much stricter conditions, i.e. $r \geqslant 0.13$ with $\alpha = 0.001$ and $\beta = 0.01$ (minimum N = 1849; see Hulley et al., 2013, p. 79). The sample is thus suitable to detect even small effects with great certainty.

3.1.1. Analysis

Preference for music features was calculated by multiplying the stated preference of each song with the mean feature rating for that song, and dividing the sum for each feature through the sum of all the preference ratings. The following formula gives an example to calculate the music preference for the *sad* feature for one participant (see also Greenberg et al., 2016):

$$\text{Preference for sad} = \frac{\text{Preference for excerpt 1} \cdot \text{Mean sad rating for excerpt 1} + \text{Preference for excerpt 2} \cdot \text{Mean sad rating for excerpt 2} + \ldots}{\text{Sum of all preferences}}$$

The music feature preference was calculated for each feature in both rating methods, i.e. for both the human-rated features and the computer-extracted music features. We then subjected the measured music feature preferences from both methods to principal component analyses to reduce the data into a smaller number of components.

To cross-validate the results from the PCA, we split the sample for version A randomly into two halves, calculated a PCA for both halves and then evaluated replicability of the component structure using target rotation (i.e. Procrustes analysis). Further, we calculated a PCA for the B version and examined fit of the component structure to the loadings extracted from version A, thus confirming our results on a different set of songs and participants.

The reported results from the principal component analysis are based on the first random half of the A version.

3.2. Results

Unless otherwise noted, all correlations reported in the results section are significant with at least $p < .01$.

3.2.1. Internal consistency

The raters showed an overall good internal consistency with a mean $\alpha = 0.80$ for version A and $\alpha = 0.79$ for version B. Internal consistencies ranged from .71 (*Depressing*) to .87 (*Systematic*) in version A, and from .72 (*Relaxing* and *Gentle*) to .87 (*Systematic*) in version B.

3.2.2. Component structure of measured music preference using human feature ratings

We subjected the calculated music feature preference with human ratings to a principal component analysis. We included the 25 music feature items from Greenberg et al. (2015). Seven items were removed due to high inter-item correlations of $r > 0.97$, namely *Tender*, *Sophisticated*, *Fun*, *Animated*, *Manic*, *Poetic*, and *Thrilling*. The remaining correlation matrix was smoothed by eigenvalue rescaling. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was 0.81, and Bartlett's test of sphericity was significant ($\chi^2(153)$=20109.90 p <.001). Variance analysis suggested three major components. We decided a priori to retain only items with a component loading of more than .50, as well as a minimum loading difference of .20 between components. We excluded the item *Party music* because it did not fit these criteria. The loading matrix was oblimin-rotated.

The first component explained 38% of the variance, the second component 31%, and the third component 24%. Overall, the model explained 93% of the variance. Inspection of the component loadings suggested we found the three established major components *Arousal* (with high loadings on features such as *Tense* and *Strong*), *Valence* (*Amusing*, not *Depressing*), and *Depth* (*Complex*, *Deep*, *Intelligent*). The component loadings can be found in Table 3. The oblimin rotation led to two item loadings above 1.0 (Tense and Warm), indicating inter-factor correlations, which were later examined.

3.2.3. Component structure of measured music preference using computer-extracted music features

Again, we subjected the measured music feature preference to a Principal Component Analysis. Since our goal was to compare the results with those of the human-rater method, we decided to exclude genre classifiers from the analysis, and thus included only mood- and sound-related classifiers. We excluded two classifiers (*Aggressive* and *Sad*) because they showed high correlations with other classifiers. Again, the correlation matrix was smoothed by eigenvalue scaling. The KMO was 0.81, and Bartlett's test of sphericity was significant ($\chi^2(171) = 21170.82$, p <.001). Variance analysis again suggested a three component solution. On terms of the same a priori criteria as in the human-rated PCA, we excluded the two classifiers *Gender* and *Relaxed* from analysis. The loading matrix was oblimin-rotated.

The first component explained 37% of the variance, the second component 26%, and the third component 18%. Overall the model explained 80 % of the variance. The first component had loadings such as *Acoustic* (r = −0.86), *Party* (r = 0.71), *Average Loudness* (r = −0.96), *Mood cluster 1 (Rousing, Passionate)* (r = 0.75), and *Mood cluster 3 (Literate, Brooding)* (r = −0.81). The second component showed loadings with *Mood cluster 2 (Fun, Cheerful)* (r = 0.84), *Mood cluster 4 (Humorous, Witty)* (r = 0.85), and *Mood cluster 5 (Aggressive, Intense)* (r −0.75), as well as with *Happy* (r = 0.78) and *Bright timbre* (r = 0.88). Lastly, the third component showed

Table 3
Component loadings for measured music feature preference using human-rated music excerpts.

Music feature	Arousal	Valence	Depth
Cerebral			
Complex	0.32	0.44	**0.90**
Deep	−0.32	−0.32	**0.79**
Intelligent	−0.31	0.16	**0.84**
Reflective	**−0.66**	−0.34	0.48
Energy			
Danceable	−0.22	**0.84**	−0.06
Emotional	0.02	−0.44	**0.83**
Gentle	**−0.89**	−0.02	0.25
Lively	0.49	**0.83**	0.02
Relaxing	**−0.83**	−0.06	0.35
Negative affect			
Depressing	0.37	**−0.90**	0.00
Sad	−0.23	**−0.88**	0.21
Tense	**1.01**	−0.16	0.13
Positive affect			
Amusing	−0.01	**0.93**	0.11
Joyful	−0.50	**0.79**	0.14
Sensual	**−0.64**	−0.02	0.42
Strong	**1.04**	0.04	0.29
Warm	**−0.82**	0.36	0.24

Note. The component loadings were calculated on base of stated preference for 10 music excerpts ("version A"). The loadings of the *Arousal* component been reversed. The loadings have been oblimin-rotated. Primary component loadings are presented in bold typeface.

loadings with *Instrumental* (r = 0.93), *Speed* (r = −0.84), *Dissonance* (r = −0.62), and *Tonal* (r = −0.59). We thus concluded that we found the same three components as in the human PCA and named them *Arousal*, *Valence*, and *Depth*, respectively. The component loadings can be found in Table 4.

Table 4
Component loadings for measured music feature preference using computer-extracted music features.

Classifier	Arousal	Valence	Depth
Cluster (Mirex)[1]			
1 (Rousing, Passionate)	**0.75**	−0.43	−0.07
2 (Fun, Cheerful)	−0.11	**0.84**	−0.12
3 (Literate, Brooding)	**−0.81**	0.03	0.40
4 (Humorous, Witty)	0.31	**0.85**	−0.20
5 (Aggressive, Intense)	0.42	**−0.75**	−0.18
Mood			
Happy	0.30	**0.78**	0.00
Party	**0.71**	−0.31	−0.35
Sound			
Acoustic	**−0.86**	0.20	0.12
Average loudness	**0.96**	0.13	0.17
Bright timbre	−0.22	**0.88**	0.09
Danceable	**0.73**	0.48	0.13
Dissonance	0.39	−0.36	**−0.62**
Dynamic complexity	**−0.85**	−0.23	0.04
Electronic	**0.89**	−0.15	0.38
Instrumental	0.19	−0.02	**0.93**
Speed (BPM)	−0.10	0.11	**−0.84**
Tonal	0.26	0.30	**−0.59**

Note. The component loadings were calculated on base of stated preference for 10 music excerpts ("version A"). The loadings have been oblimin-rotated. Primary component loadings are presented in bold typeface.
(1) Mood clusters (MIREX) (Hu & Downie, 2007):
Cluster 1: Rowdy, Rousing, Confident, Boisterous, Passionate.
Cluster 2: Amiable/Good natured, Sweet, Fun, Rollicking, Cheerful.
Cluster 3: Literate, Wistful, Bittersweet, Autumnal, Brooding, Poignant.
Cluster 4: Witty, Humorous, Whimsical, Wry, Campy, Quirky, Silly.
Cluster 5: Volatile, Fiery, Visceral, Aggressive, Tense/anxious, Intense.

3.2.4. Correlation between component preferences

Next, we tried to determine the equality of both component structures. Usually, a factorial invariance test would be the method of choice; however, such tests require that both methods measure the exact same variables. While we did found a good accordance of both rating methods, the assessed features are not the same and thus cannot be used in a factorial invariance test.

Instead, we used Pearson correlations to examine the inter-component correlations within and between both methods of measurement. Within the human ratings, only the correlation of *Depth* and *Arousal* became significant with r = −0.30 (p < 0.001). The ESSENTIA-extracted music preference components showed relatively small intra-method correlations between r = 0.14 and r = −0.19 (all *n.s.*), as seen in Table 5.

Inter-method correlations showed that the largest correlations between the components were usually not found between the matching component pairs, but in a negative relationship with other components. For instance, Human Arousal was mostly negatively associated with ESSENTIA Valence (r = −0.83), and the loading with ESSENTIA Arousal was also found, but smaller (r = 0.52). The same is true for Human Depth, which correlated negatively with ESSENTIA Arousal (r = −0.73), and to a lower degree with ESSENTIA Depth (r = .60). Human Valence showed negative relationships with ESSENTIA Arousal and Depth (r = −0.47, and r = −0.38, respectively), but only a low and non-significant correlation with ESSENTIA Valence (r = 0.18, *n.s.*). These results suggest that, while they do capture some of the same features, the music preference components resulting from human ratings and machine-extracted features do not align directly with each other.

3.2.5. Confirmation of component structure

We calculated a PCA for the second half of the dataset from version A. We then subjected the component structure to a Procrustes analysis with target rotation to examine the similarity of both results. We prefer Procrustes analysis over confirmatory factor analysis (CFA) because the obtained three component solution from the first half of the sample showed manifold secondary loadings of the music ratings. For instance, and possibly by definition, the perception of complexity has a substantive negative loading of .31 on the first component (*Arousal*) and also on the second component (*Valence*) of .44, with a high main loading of .90 on the third component (*Depth*). This holds true for most of the items. Whereas CFA is best suited to the analysis of simple structure models (Loehlin, 1998), CFA is inappropriate when cross-ladings on multiple factors occur (McCrae, Zonderman, Costa, Bond, & Paunonen, 1996; Hopwood & Donnellan, 2010), as is the case in our data. The consequence of failing to include cross-loadings in a CFA model is that they are then assumed to be zero and any true deviation from zero contributes to model misfit.

First we examined the two random halves of version A of our survey. Comparison of the human-rated music preference component structure revealed a root mean squared error (RMSE) of .01 and a between-method correlation of 1.00 (p < .001), indicating (almost) perfect fit of the loading matrices. The ESSENTIA-rated music preference component structure also shows an excellent fit, with a RMSE of .02 and a correlation of 1.00 (p < .001).

Comparison of the component loadings found in version B with those of version A showed a great fit for the human-rated method: The RMSE was still low with .05, and the between-model correlation was r = 0.98 (p < .001). The ESSENTIA-rated method showed a lower, yet still good model fit with RMSE = 0.17 and a correlation of r = 0.71 (p < .001). The results suggest that the two different sets of music pieces in version A and B lead to the same principal component structures. Further results can be obtained from Table 6.

Table 5
Within- and between-method component score correlations.

	1	2	3	4	5
1 Human Arousal	–				
2 Human Valence	−0.01[a]	–			
3 Human Depth	**−0.30**	−0.02[a]	–		
4 ESSENTIA Arousal	**0.52**	**−0.47**	**−0.73**	–	
5 ESSENTIA Valence	**−0.83**	0.18	−0.16	−0.19	–
6 ESSENTIA Depth	**−0.49**	**−0.38**	**0.60**	−0.18	0.14

Note. N = 786. All correlations are significant with p < .001, except (a) *n.s.*. Correlations larger than 0.30 are presented in bold typeface.

Table 6
Results of symmetric procrustes analysis.

	SS	RMSE	MPE	Correlation	p
A first half vs. A second half					
Human	0.00	0.01	0.01	1.00	<.001
ESSENTIA	0.01	0.02	0.02	1.00	<.001
B vs. A first half					
Human	0.05	0.05	0.16	0.98	<.001
ESSENTIA	0.50	0.17	0.15	0.71	<.001

Note. SS = Sum of Squares, RMSE = Root Mean Squared Error, MPE = Median procrustes error.

3.3. Discussion

In the second study, we examined the structure of music preference as assessed using human ratings and computer-extracted features in a large online sample. The principal component analyses showed the same components in both methods of feature ratings. The components (*Arousal*, *Valence*, and *Depth*) match those found in previous research using excerpt-based assessment (Greenberg et al., 2016) and self-report (Fricke & Herzberg, 2017). We do have to note that the components are less distinct in the computer-extracted feature method, especially the *Depth* component. We attribute this to the fewer number of features, and the fact that we're not able to freely select extracted music features, but instead have to use those offered by the analysis software. Also, the music classifiers originate from various different studies, often with their own ground truth data of varying quality. It might be worthwhile to replicate or re-train the classifiers on a single, well-constructed ground truth dataset. Further, it could be interesting to examine if more mood and sound classifiers could be learned from human-rated music pieces.

The cross-method correlations indicated that some of the computer-extracted components were more characterized by adverse relationships with other human components than by the correlation with the matching component. We suppose this could be due to the fact that the methods captured different facets of the same construct. For instance, the human *Arousal* component was characterized by loadings on *Tense* and *Strong* features, while the ESSENTIA *Arousal* component was characterized through *Loud*, but *Danceable*, mostly *Electronic* music. ESSENTIA's *Depth* component solely caught the sound-related features, such as *Instrumental* and low *Speed*, while the human-rated *Depth* component captured features such as *Complex*, *Intelligent*, and *Emotional*. Importantly, these results suggest that the two measurements of music preference cannot be used interchangeably; i.e., while both methods capture some of the same features and show similarities, they are not directly aligned and differ regarding their contents. The similarities between methods are also limited to two of the three components – the *Valence* components only correlate with r = 0.18.

Still, finding the three-component AVD structure using two different rating methods proves the component structure robust and paves the way for large-scale music analysis to measure music preference as a psychological phenomenon. There still are satisfactory correlations between two of the components of the two methods, which suggests that insights from previous research can potentially be applied to computer-extracted music feature preference (these insights should, however, be formally replicated). Also, we found almost identical model fit for two random samples who rated their preference for the same songs (version A), and a great model fit for results obtained from a different set of participants and music excerpts (version B). These results indicate that the extracted components are indeed robust both within and across methods.

4. General discussion

Our research yielded two main results: First, human-rated and computer-extracted music features are mostly aligned and measure similar concepts of music attributes. And second, measured music feature preference can be described on the three components *Arousal*, *Valence*, and *Depth* – regardless whether the features were rated by humans or extracted by the ESSENTIA software library. Thus, we showed that music features extracted with the ESSENTIA software library can be used to assess individual music preference.

The correlations of human music feature ratings and computer-extracted features suggest that both methods measure similar concepts of music features. However, it remains unclear whether machines are able to identify more detailed music features. One of the main problems with examining this would be the detailed annotation of music features for a large body of music to generate the ground truth data for the SVMs. However, our research indicated that most music feature ratings are relatively robust across different rater groups. Therefore, the rating effort could be crowdsourced, with raters rating only a handful of songs and features at a time and integrating these results to a combined dataset.

The second study found the three-component structure Arousal, Valence, and Depth (AVD) for both human ratings and computer-extracted features. The results from the two methods are, however, not identical: Most of ESSENTIA's music preference components are more characterized by adverse relationships with other Human preference components, rather than their matching counterparts. Only ESSENTIA's Depth factor shows the greatest inter-component correlation with the respective matching human

component (which itself correlates more with negative ESSENTIA Arousal). As discussed in Study 2, this could be due to the components capturing different facets of the same underlying factors. Another reason could be the selection of attributes as a whole: We had to rely on the available music classifiers from ESSENTIA and could not select our own adjectives. We might find a more aligned component structure if we had access to more detailed and nuanced classifiers, as discussed above.

The implication from this finding is mainly that we have to choose our assessment method carefully concerning the questions we're trying to answer. Intuitively, human raters should be able to more precisely distinguish between music characteristics. However, they are also prone to biases, such as their own music preference. Computer-extracted features are especially useful when there is much data to rate, as it can be done automatically, and when no human rating data is available. Also, they can avoid objectivity restraints arising from human biases. Computer-based feature extraction thus should be especially useful when examining actual listening data, such as in the construction of music recommendation systems. We also think that music preference measured from computer-extracted features is accurate enough to examine relationships with other variables, such as personality dimensions, in large samples. However, until a higher similarity can be obtained between computer-extracted features and human ratings, human ratings are more descriptive and accurate for single songs, and should also yield a more detailed assessment of individual music preference.

The research on music feature preference showed the AVD structure using various methods: In excerpt-based assessment (Greenberg et al., 2016), in self-report assessment (Fricke & Herzberg, 2017), in excerpt-based assessment with another group of raters (Study 2) and in excerpt-based assessment using computer-extracted features (Study 2). It is thus safe to say that music feature preference has proven to be a robust personality phenomenon across different measures.

With yet another method for the assessment of music feature preference, the question arises how measurements from this method relate to established research on the relationship of music preference and personality. While a recent meta-analysis found that the correlations of musical styles (i.e. the MUSIC-Model) with the Big Five personality dimensions are rather weak (Schäfer & Mehlhorn, 2017), it would be interesting to examine those relationships more deeply with music *feature* preference. In self-report, relationships between Openness, Neuroticism and Depth, as well as between Extraversion, Agreeableness and Valence were found (Fricke & Herzberg, 2017). It would be worthwhile to see if these relationships are robust and found in excerpt-based assessment, as well, and more so in computer-based assessment, as this method is suitable to be applied on a large scale by music content providers such as Spotify, Deezer, or Apple Music. Also, since music features stem from features that can be used to describe various other constructs, relating preference for certain music features to those other areas, such as creative arts or movie preference, could be interesting for research and business appliances. Lastly, we'd like to encourage the examination of relationships with personality constructs other than the Big Five. For instance, research on the relationships with *Sensation Seeking* (e.g. Litle & Zuckerman, 1986) could be relived with music feature preference, or research could examine relationships with other constructs that are likely related to musical preference, such as *Need for Uniqueness* or *Hypersensitivity*.

Our results also demonstrate the accuracy of modern music analysis libraries and classifiers. With these tools on hand, it is possible to measure music feature preference from other music sources, such as actual listening behavior, and use insights from previous research regarding the psychological phenomenon of music feature preference. There have been some efforts to include lyric analysis in music information retrieval appliances, and even mood classification (e.g. Van Zaanen & Kanters, 2010). Integrating or comparing these methods with music feature analysis could open up yet another way to assess music preference, and for various other research questions.

Computer-based music analysis primarily enables researchers to examine a large pool of music pieces in short time. Using this technology, we can infer insights about music preference not only on an intra-personal, but on a national or even global level. For instance, by examining music charts (e.g. the Billboard Hot 100) over a large timespan, researchers could see if national music preference changed over the years, and relate these insights to similar research, like the development of intra-individual music preference over the lifespan (see Bonneville-Roussy et al., 2013). Another interesting perspective is to compare the features of the music charts of different countries, thus examining cultural differences in music perception and preference. Current research mostly focused on western music. Especially the sound-related classifiers offer a way to objectively assess music features, enabling unbiased comparability between different cultures. Since our current research examined mostly western participants, such research could also examine the generalizability of our findings. Lastly, computer-based music feature analysis can be used to infer individual music preference from actual listening behavior. It would be interesting to see if music preference assessed that way show the same component structure, and to see if they have any predictive power for biographic variables, such as age, gender, and country of origin, or even personality traits, such as the Big Five or Sensation Seeking.

Music feature preference is a robust personality phenomenon. We have various tools on hand to assess it, and many research questions to answer. Lots of potential use-cases emerge from the result that computers can actually measure music preference. In this paper, we further advanced knowledge of how music preference is structured and paved the way for many more research and real-world applications utilizing the assessment of individual music preference.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jrp.2018.06.004.

References

Bogdanov, D. (2013). *From music similarity to music recommendation: Computational approaches based on audio features and metadata* Ph. D. dissertation. Barcelona (Spain): Universitat Pompeu Fabra.

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., & Mayor, O. (2013). ESSENTIA: An audio analysis library for music information retrieval. In *International Society for Music Information Retrieval Conference (ISMIR'13)* (pp. 493–498).

Bonneville-Roussy, A., Rentfrow, P. J., Xu, M. K., & Potter, J. (2013). Music through the ages: Trends in musical engagement and preferences from adolescence through middle adulthood. *Journal of Personality and Social Psychology, 105*(4), 703–717.

Doi, H., Basadonne, I., Venuti, P., & Shinohara, K. (2018). Negative correlation between salivary testosterone concentration and preference for sophisticated music in males. *Personality and Individual Differences, 125*, 106–111.

Downie, J. S. (2003). Music information retrieval. *Annual Review of Information Science and Technology, 37*(1), 295–340.

Fricke, K. R., & Herzberg, P. Y. (2017). Personality and self-reported preference for music attributes. *Journal of Research in Personality, 68*, 114–123.

Greenberg, D. M., Baron-Cohen, S., Stillwell, D. J., Kosinski, M., & Rentfrow, P. J. (2015). Musical preferences are linked to cognitive styles. *PloS One, 10*(7), e0131151.

Greenberg, D. M., Kosinski, M., Stillwell, D. J., Monteiro, B. L., Levitin, D. J., & Rentfrow, P. J. (2016). The song is you. *Social Psychological and Personality Science, 7*(6), 597–605.

Hopwood, C. J., & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review, 14*(3), 332–346.
Hu, X., & Downie, J. S. (2007). Exploring mood metadata: Relationships with genre, artist and usage metadata. In *ISMIR* (pp. 67–72).
Hulley, S. B., Cummings, S. R., Browner, W. S., Grady, D. G., & Newman, T. B. (2013). *Designing clinical research*. Lippincott Williams & Wilkins.
Langmeyer, A., Guglhör-Rudan, A., & Tarnai, C. (2012). What do music preferences reveal about personality? *Journal of Individual Differences, 33*(2), 119–130.
Litle, P., & Zuckerman, M. (1986). Sensation seeking and music preferences. *Personality and Individual Differences, 7*(4), 575–578.
Loehlin, J. C. (1998). *Latent variable models: An introduction to factor, path, and structural analysis*. Lawrence Erlbaum Associates Publishers.
McCrae, R. R., Zonderman, A. B., Costa, P. T., Jr., Bond, M. H., & Paunonen, S. V. (1996). Evaluating replicability of factors in the revised neo personality inventory: Confirmatory factor analysis versus procrustes rotation. *Journal of Personality and Social Psychology, 70*(3), 552–566.
McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., & Battenberg, E. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in science conference* (pp. 18–25).
McKay, C. (2010). *Automatic music classification with jMIR*. McGill University.
North, A. C., & Hargreaves, D. J. (1999). Music and adolescent identity. *Music Education Research, 1*(1), 75–92.
Rentfrow, P. J., Goldberg, L. R., & Levitin, D. J. (2011). The structure of musical preferences: A five-factor model. *Journal of Personality and Social Psychology, 100*(6), 1139–1157.
Rentfrow, P. J., Goldberg, L. R., Stillwell, D. J., Kosinski, M., Gosling, S. D., & Levitin, D. J. (2012). The song remains the same: A replication and extension of the MUSIC model. *Music Perception, 30*(2), 161–185.
Rentfrow, P. J., & Gosling, S. D. (2003). The do re mi's of everyday life: the structure and personality correlates of music preferences. *Journal of Personality and Social Psychology, 84*(6), 1236–1254.
Schäfer, T., & Mehlhorn, C. (2017). Can personality traits predict musical style preferences? a meta-analysis. *Personality and Individual Differences, 116*, 265–273.
Van Zaanen, M., & Kanters, P. (2010). Automatic mood classification using TF* IDF based on lyrics. In *ISMIR* (pp. 75–80).
Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences, 112*(4), 1036–1040.
Zweigenhaft, R. L. (2008). A do re mi encore: A closer look at the personality correlates of music preferences. *Journal of Individual Differences, 29*(1), 45–55.