Energy, Typicality, and Music Sales: A Computerized Analysis of 143,535 Pieces

Adrian C. North, Amanda E. Krause, Lorraine P. Sheridan, and David Ritchie

Abstract

Research on musical preference has been dominated by two approaches emphasizing respectively the arousal-evoking qualities of a piece or its typicality of the individual's overall musical experience. There is a dearth of evidence concerning whether either can explain preference in conditions of high ecological validity. To address this, the present research investigated the association between sales of 143,353 pieces of music, representing all that music that has enjoyed any degree of commercial success in the United Kingdom, and measures of both the energy of each piece (as a proxy for arousal) and the extent to which each piece was typical of the corpus. The relationship concerning popularity and energy was U-shaped, which can be reconciled with earlier findings, and there was a positive relationship between the typicality of the pieces and the amount of time they featured on sales charts. The population-level popularity of an entire corpus of music across several decades can be predicted by existing aesthetic theories, albeit with modifications to account for market conditions.

Key words: Music, sales, arousal, energy, typicality

Energy, Typicality, and Music Sales: A Computerized Analysis of 143,535 Pieces

Experimental aesthetics is one of the oldest fields in psychology (see Fechner, 1876), and research methods have been influenced heavily since by the experimental approach via which the field initially attracted significant attention. In the context of music, this typical methodology involves playing a short 'stimulus' to lab-based undergraduate participants which is often monophonic, short (i.e., typically lasting less than a minute), and regularly composed specifically for the research in question according to a statistical or similar rule. Participants then provide verbal measures of liking for the music or researchers collect cognitive and/or physiological response data, such as skin conductance. This approach clearly allows fine-grained experimental control, but has attracted considerable criticism in recent years for its lack of ecological validity, as it reflects neither 'real music' as experienced by the majority of the population, 'real participants' (given the well-documented problems associated with employing student research samples, e.g., Henrich, Heine, & Norenzayan, 2010; Wintre, North, & Sugar, 2001), or 'real responses', which involve much more complex behaviours, such as radio station selection or music purchasing. In contrast, the present research tests whether the two major theories of musical preference developed via lab-based methods over recent decades can predict variations in the commercial popularity of 143,353 pieces for which sales data exists, which in effect represent the entire corpus of music that has enjoyed any degree of commercial note in the United Kingdom. The present findings demonstrate that variations in the degree of commercial success of this entire corpus of music can be accommodated by the two theories in question, but that both require modification in order to reflect the lack of ecological validity in the methods by which they have been developed.

Arguably the most influential theory in attempts to explain music preference is Berlyne's (1971) psychobiological approach. This states that preference for music (and other art forms) is related to the degree of arousal they evoke in the ascending reticular activating system (ARAS). The theory argues that people should like most that music which is located towards the center of the

'arousal potential' continuum, with levels of liking showing a progressive decline towards either end of this continuum, leading to a so-called 'inverted-U' relationship between liking for music and its arousal potential. Berlyne argued that the variables that cause music to bring about arousal fall into three categories. Psychophysical variables represent the intrinsic physical properties of the music, such as its tempo and loudness; ecological variables refer to the signal value of the music, such as the associations one might draw between a significant life event and a particular piece; and collative variables refer to the informational properties of the music, such as its degree of complexity (e.g., the degree of variation in the melody or rhythm), redundancy (e.g., the degree of repetition within the piece), and familiarity (which reflects the amount of new information that a piece provides to the listener in a given listening episode). Berlyne (1971, p.69) claimed that these collative variables are "the most significant of all for aesthetics".

Although the portion of the theory concerning the ARAS is undoubtedly contentious from a physiological standpoint (see e.g., Martindale, 2007), the apparent seductiveness of Berlyne's approach for researchers was no doubt aided by its consistency with the aesthetic findings of Fechner (1876) and also the theories of the classical Greek philosophers. Fechner argued for the importance of the aesthetic mean, or the notion that beauty is associated with an absence of extremes. Similarly, as Berlyne (1971, p.123) noted, "Plato (in the Statesman) wrote that all arts are 'on the watch against excess and deficit ... [in that] the excellence and beauty of every work of art is due to this observance and ... a standard removed from the extremes.' A little later, Aristotle (in the Nicomachean Ethics) made the same point with exemplary conciseness: 'A master of any art avoids excess and defect but seeks the intermediate and chooses this.'" However, perhaps the clearest reason for the popularity of Berlyne's approach in the literature is the number of studies that have used information theory and/or physiological measures to conceptualize music and which have provided findings consistent with the theory (see e.g., Crozier, 1974; McMullen, 1974; McMullen & Arnold, 1976; Simon & Wohlwill, 1968; Vitz, 1966; and reviews by Hargreaves, 1986; North & Hargreaves, 2008). Of greatest relevance to the present research, Simonton (1980) analyzed the

initial six notes of 15,618 classical music themes for their melodic originality - a measure of the statistical improbability of the transitions between notes, which is clearly analogous to Berlyne's notion of complexity - showing that this had an inverted-J relationship with ecologically-valid measures of popularity.

However, in addition to the precise details of its physiological basis, Berlyne's theory and approach can be criticized on two other significant grounds. The first of these concerns the ecological validity of both the music employed and participants' responses to this. With regard to the music, careful reading of the methodology and appendices of the published research indicates that, in order to produce the tails of the inverted-U distribution of the liking-arousal relationship, researchers have been forced to employ musical stimuli with such extreme levels of complexity, redundancy, etc. that it is possible to question whether they are truly 'musical' in the sense that the public would recognize. For example, North and Hargreaves (1995) were able to identify an inverted-U relationship between ratings of liking and complexity assigned by laboratory participants to excerpts of new age music: but listening to some of the pieces located towards the poles of the complexity dimension (e.g., *Courage* by Jon Hassell, *Idle chatter* by Paul Lansky, or *Sequence symbols* by James Dashow) suggests that the extremely repetitive or complex nature of the high art music concerned means that it is some distance from what the general population might regard on a day-to-day level as representing 'music'. The extreme levels of complexity of these pieces and their lack of widespread popularity is understandable, but they also sound very little like the music usually played on the radio or in concert halls to which the majority of the population is exposed on a regular basis. Such a problem is common to many of the other studies cited above in support of Berlyne's theory. More simply, while Berlyne's theory appears to do well in distinguishing very unpopular from reasonably popular music, it is by no means clear how well the theory performs when attempting to distinguish the varying popularity levels of what most would regard as the 'normal' music that is played on the radio or regularly streamed online. Can the theory

distinguish a number 1 hit single from a neglected album track that was in the top 40 country sales chart for only a week?

A second significant problem with Berlyne's approach concerns participants' responses. When a lab-based participant produces a skin conductance reading or assigns a rating on a scale concerning liking for a particular piece it is not clear that these measures are truly analogous to a response with much greater ecological validity or practical relevance, such as music purchasing or radio station selection. Indeed, given these two limitations, it is interesting that other research using small samples of 'real' music of more typical levels of arousal potential and more naturalistic measures of liking for music has failed to support Berlyne's theory. For example, Russell (1987) found that the appearance of a song in music sales charts was unsurprisingly associated with it possessing greater familiarity, but that this change in familiarity was not associated with changes in liking for the song in question, contrary to Berlyne's theory. Similarly, North and Hargreaves (2000a) found that, when asked to select music to listen to while exercising, participants chose that which would further increase rather than moderate their level of arousal; and when asked to select music to listen to while relaxing, participants chose that which would further decrease rather than moderate their level of arousal. There is a dearth of research utilizing naturalistic responses to 'real' music, and what little there is falls some way short of supporting Berlyne's theory.

In addition to this, some researchers have argued that the relationship between liking for music and the variables considered by Berlyne is not of the nature described by the latter, and that different underlying mechanisms drive this relationship. Best-known among these is Zajonc's (1968) assertion that, "mere repeated exposure of the individual to a stimulus is a sufficient condition for the enhancement of his attitude toward it" (p.1), contrary to Berlyne's arguments about the relationship between liking and familiarity; and his later claim (Zajonc, 1980) that such a process can occur without any reference to cognition. Several authors (e.g., Bornstein & D'Agostino, 1994; Seamon, Brody, & Kauff, 1983) have argued that the mere exposure affects liking because it increases perceptual fluency, or the ease with which a given stimulus can be

processed. While the validity of Zajonc's approach is a separate topic, the important point for the present research is that it illustrates that Berlyne's theory of the relationship (and underlying mechanism) between positive responses and variables such as familiarity is by no means supported universally.

Martindale, Moore, and West (1988) among others built on Berlyne's theory, the mere exposure hypothesis, and the notion of processing fluency to argue that the variables considered by most of the research on Berlyne's theory are not the most relevant to aesthetics. They argued that preference is determined by the extent to which a particular artwork is typical of those in its class (rather than the amount of arousal it evokes), and explanations of this have tended to invoke connectionist models (rather than psychobiology). This approach claims that preference is positively related to typicality because typical stimuli give rise to stronger activation of the salient cognitive categories. A consistent conclusion of this research, much of which was conducted in the 1980s and 1990s, has been that not only can typicality explain preference, but also that it can explain a much greater portion of the variance in this than can Berlyne's arousal-based approach. For example, Martindale and Moore (1989) reported that complexity accounted for 4% of the variance in liking for classical music themes, whereas the typicality of the themes accounted for 51% of the variance. Similar results obtained using stimuli other than music are reported by Hekkert and van Wieringen (1990), Martindale, Moore, and Borkum (1990), Moore and Martindale (1983), Whitfield (1983), and Whitfield and Slatter (1979). Martindale, et al. (1988, p. 94) argue that results such as those described here, "suggest that collative variables are probably a good deal less important in determining preference than Berlyne thought them to be. Furthermore, they probably determine preference via mechanisms different than those proposed by Berlyne".

The typicality approach might on the surface appear inconsistent with, or simply better than, Berlyne's theory, such that the latter ought perhaps be discarded as a less effective predictor of preference. However, North and Hargreaves (2000b) argued that typicality at least incorporates (perhaps to a considerable extent) the arousal-evoking qualities of the music in question, such that

conceptions of music in terms of the latter still had the potential to be useful. Evidence concerning this latter point is lacking, even though it is possible to test the claim by investigating the correlation between arousal and typicality.

It is also possible to make a second, more intuitive objection to the proposed relationship between liking and typicality in the context of the commercial reality of the music industry. Given the crowded market, a piece of pop music in particular might be more successful if it is innovative and therefore distinctive relative to its commercial competitors, such that popularity is not positively and monotonically related to typicality. A similar point is made by Martindale (1990), who argued that a need to maintain people's attention explains the tendency he observed in many art forms to progress over time towards generating ever greater levels of arousal in the audience.

A third possible problem for the typicality-based approach is that again the research has tended to focus on small numbers of artistic stimuli of sometimes limited ecological validity. This is attributable to some extent to a practical issue of computing power: in order to consider the extent to which a given piece of music is 'typical' requires consideration of the corpus as a whole, and all but the most powerful of the computers of the past two decades have been unable to cope with such large datasets. A similar point can be made concerning existing tests of Berlyne's theory, which have also been unable (in practice if not in theory) to encode very large datasets of music and use these to provide more definitive tests.

Recent advances in computing power make it possible to test these issues, however. As such, the present research employed an initial database of over 35 million pieces of music, which contained all music that has obtained a commercial release via one of over 400,000 record labels in Europe, North America, and Australasia. For the purposes of the present research, this larger database was filtered to only and all of those 143,353 pieces for which data existed concerning United Kingdom sales. More simply, the research employed all that music that had enjoyed any degree of commercial success in that country. Data was collated for each piece on several variables. The first of these, energy, addressed the arousal-evoking qualities of each piece. A second variable,

'general hit popularity' (and a corresponding United Kingdom-only version of this variable), used chart positions in various United Kingdom and United States sales charts for each piece over time to address the popularity of each piece at the population level. A similar variable 'general hit appearance' (and a corresponding United Kingdom-only version of this variable) considered each piece in terms of simply the number of weeks spent on United Kingdom and United States sales charts. The database also contained information for each piece in terms of beats per minute (BPM) and six specific emotional connotations: these data (and energy) were used to operationalize typicality by computing, for each piece, the difference between that piece and the mean value for the corpus as a whole (and transforming this score to remove the direction of any difference). The research tested three hypotheses concerning these variables, namely;

H1. Given the predictions of Berlyne's theory, there should be an inverted-U relationship between energy and both hit popularity and hit appearance, indicating that moderately-arousing music is liked most.

H2. Given the predictions of the typicality approach, there should be a negative linear relationship between difference scores and both hit popularity and hit appearance, indicating that typical music is liked most.

H2a. The relationship between popularity and typicality may be different in the case of pop music, indicating the greater commercial success of less typical music and a greater market tolerance of innovation.

Method

Dataset

The research employed an adapted version of a master dataset used extensively within the music industry, with the adaptation created in partnership with a private sector organization. The master database contains information on over 38 million pieces of recorded music, which in effect represents all music recordings ever released on a commercial basis in Europe, North America, and Australasia since the beginning of the 20th century (including recordings of pieces composed before this date). The master database is compiled by a company, which aggregates information globally from over 400,000 record labels. The master database represents the canonical music catalogue used by radio stations, recording companies, and other media in music programming and other similar activities. On entry into the master dataset, the company concerned classifies each piece into one of 23 genres (namely, alternative/indie, blues, cast recordings/cabaret, children's, Christian/gospel, classical/opera, comedy/spoken word, country, electronica/dance, folk, instrumental, jazz, Latin, new age, pop, rap/hip hop, reggae/ska, rock, seasonal, soul/R&B, soundtracks, vocal, and world) on the basis of the recording artist in question: the initial classification of an artist incorporates information provided by the recording company in question. Note that tracks classified as 'comedy/spoken word' were deleted from the present dataset because the great majority did not contain any music, and any music they contain is clearly not the focus of the remainder. Pieces were also deleted for minority genres, for which there were fewer than 100 exemplars that also had popularity data. Created on 30 March 2015, the subset of this master dataset used in the present research contained 143,353 pieces of music, which were selected as those for which data also existed concerning sales in the United Kingdom, such that the pieces employed were all and only those that had enjoyed any commercial success whatsoever in that country: they represent a complete commercial musical culture.

Energy. The energy value for each piece was calculated via an algorithmic process that produced a score for each in turn based on its specific features: this approach is preferable to assigning scores to individual tracks on the basis of meta-data, such as genre classification, as it directly addresses the characteristics of the piece in question. The first step was establishing a set of

training tracks, consisting of 100 exemplar ‘calm' and 100 exemplar ‘energetic’ pieces, which were selected by a team comprising two students who were heavy music consumers, a musicologist, and an audio engineer working collaboratively. This set of training tracks was used in order to train an AI process (detailed in U.S. Patent No. 20100250471, 2010; and U.S. Patent No. 20080021851, 2008) about the sonic differences between energetic and calm tracks using mathematical vectors based on the combinations of 11 sound properties (e.g., tempo, beat, pitch, and rhythm). Via this AI process, the computer compared each individual exemplar track against the remaining 99 using an algorithm: if in the 10 most acoustically-similar tracks (again defined according to 11 computer-analyzed sound properties such as tempo, beat, pitch, and rhythm) there was a majority from the same proposed class as the seed track (i.e., calm versus energetic) then the target piece was regarded as having been classified appropriately. The initial batch of tracks yielded a successful classification rate of 92%, and the 18 incorrectly classified tracks were then replaced by others in subsequent iterations of the same process until all 200 of the seed tracks could be regarded as classified appropriately by this process. The trained AI process (detailed in U.S. Patent No. 20100250471, 2010; and U.S. Patent No. 20080021851, 2008), referred to as an ‘energy classifier’, was then used to process every track in the database, and assign an energy value to each on the basis of the degree of similarity between its own values on the 11 sound properties and the values of the training tracks. A similarity engine combined scores on 69 differing combinations of the 11 sound properties to determine the degree of similarity between a given piece and the other pieces in the database: this was accomplished by examining the degree of similarity on the values for each of the 69 combinations for each track in turn relative to the remainder of the tracks in the database. Each track was then assigned an energy value based on the similarity values so that the greater the similarity between two tracks so the greater the similarity in their energy scores: high values indicate an energetic track while low values indicate a calm track. The research team also carried out an informal human-listening test of 1000 tracks from the entire database, selected via a quasi-random process, which involved checking the face validity of relatively low, moderate, and high

energy values produced by the AI system. This non-statistical exercise corroborated that the computer scoring was producing scores that reflected human perceptual experience of the music.

Beats per minute (BPM). Initially, we tested five different algorithmic measures of BPM for each of the genres employed in the present research. These candidate algorithms were based on the industry-standard open source C++ library developed by the Music Technology Group of Pompeu Fabra University (http://essentia.upf.edu). The outputs of each algorithm were then compared against human ratings of a sub-sample of tracks from each of the genres. The two algorithms that produced outputs with the highest correlation with the human ratings were then combined and subsequently employed in the present research. The BPM value for each piece was determined via computerized measurements that were taken for each successive 30-second segment of each track to allow for *rallentando* and other forms of tempo variation within the track. The tempo values for each segment were subsequently averaged to provide a single BPM value per piece. Once values had been calculated for each track, the same informal human listening test as described under the 'Energy' sub-heading indicated that the outputs of this process have good face validity, as they provide a good overall assessment of tempo; and separate unpublished tests of the accuracy of the process (versus manual measurements of tempo) carried out prior to commencement of the current research also suggest that this approach performs well.

Hit popularity and hit appearance. A general hit popularity score was assigned to each piece based on data from the United Kingdom and United States charts, and a corresponding United Kingdom hit popularity score was also assigned that employed only United Kingdom sales chart information. The measures incorporated data from general charts as well as genre-specific and regional charts. Each chart was assigned a weighting based on the size of the region covered (e.g., a national chart was weighted heavier than a regional chart, with the extent of the difference depending on the size of the region in question); whether the chart addressed singles or albums (with singles charts weighted heavier albums charts, as they are a more direct reflection of the popularity of the specific track in question); and whether the chart was general versus genre- or

region-specific (with the extent of the difference in weighting of specific genre charts depending on the popularity of the genre and size of the region in question). For example, the United Kingdom singles chart was assigned a weighting of 1; the corresponding albums charts were assigned a weighting of .500 (i.e., 1/2); the United Kingdom classical specialist albums chart was assigned a weighting of .167 (i.e., 1/6); the United Kingdom Asian singles chart was assigned a weighting of .143 (i.e., 1/7); and the Scottish albums chart was assigned a weighting of .125 (i.e., 1/8). For each track per chart, the popularity score was calculated as 1 divided by (peak chart position multiplied by chart weighting), so that higher scores indicate greater popularity.

Each piece was also assigned two hit appearance scores, namely a United Kingdom score (again based on only United Kingdom chart data), and a general hit appearance score, utilizing data from both the United Kingdom and United States charts. 'Hit appearance' scores were calculated as simply the number of weeks the piece appeared on each of the charts (irrespective of chart position), and represent an arguably less sophisticated but more direct representation of chart performance.

Mood scores. For each track, a score was calculated for each of six mood clusters, namely mood 1 = clean, simple, relaxing, mood 2 = happy, hopeful, ambition, mood 3 = passion, romance, power, mood 4 = mystery, luxury, comfort, mood 5 = energetic, bold, outgoing, and mood 6 = calm, peace, tranquility, respectively. These moods were employed at the discretion of the music industry at the time the initial database was devised, and are regarded by the industry as most relevant to radio programming (and similar commercial uses): nonetheless, they possess good face validity as 'typical' responses to music. The mood scores were based on seed ratings of 300 pieces thought to represent a good range of all the moods concerned. Again, to begin the process of processing the scores, six musicians and sound engineers provided ratings of how the music made them feel in order to create a training set of tracks for the AI training. The development of the mood scores involved a three-step machine learning process, similar to that for the 'Energy' score (U.S. Patent No. 20100250471, 2010; U.S. Patent No. 20080021851, 2008). First, each piece was

analyzed according to audio descriptors based on melody, harmony, tempo, pitch, octave, beat, rhythm, noise, brilliance, and chord progression. Second, as per the energy score, a similarity engine combined scores on 69 differing combinations of the audio descriptors to determine the extent to which each track was similar to the others in the database. Third, each of the six mood scores for each piece were then determined on the basis of the mood scores assigned to similar tracks and the degree of similarity between those and the target piece on the 69 combinations of the audio descriptors. This allowed the computer to allocate percentage scores to each track that represented the extent to which it fitted each of the six moods. The same informal human listening test as described under the 'Energy' sub-heading indicated that the outputs of this process have good face validity.

Difference scores. The corpus mean was calculated for energy, bpm, and the six mood scores; and a difference score for each piece was then calculated as the summed difference between its own scores on those variables and the mean values for the corpus as a whole. (Note that energy was included in the difference scores given the arguments in North and Hargreaves (2000b) which identify that energy is likely a component of typicality, but nonetheless a distinct concept in its own right as it has a psychobiological basis rather than the cognitive basis of typicality.) The resulting value, if negative, was multiplied by -1 so that this total difference score serves as a measure of the typicality of the piece relative to respectively the corpus (irrespective of the direction of difference). Separate scores for each piece were also calculated within each genre.

Results

A series of curvilinear regression analyses was performed to test Hypothesis 1, namely that there should be an inverted-U relationship between popularity and energy. A separate curvilinear regression analysis considered each of general hit popularity, general hit appearance, UK hit popularity, and UK hit appearance as the dependent variable. The results of these are summarized in Table 1 showing the direction of the beta values for energy and energy-squared, and whether the

inclusion of the latter added significantly to the proportion of the variance explained by the monotonic relationship. There was a significant quadratic relationship between energy and each of general hit popularity, United Kingdom hit popularity, general hit appearance, and United Kingdom hit appearance; and we discuss the nature of these relationships in more detail shortly.

- Table 1 about here -

A series of Pearson's *r* correlations addressed Hypothesis 2, namely that there should be a negative linear relationship between the difference scores and each of the measures of popularity. The results of these are summarized in Table 2. The data were consistent with H2 in the case of general hit appearance and United Kingdom hit appearance, although the relationship was not found when popularity was defined as general hit popularity or United Kingdom hit popularity.

To test Hypothesis 2a, that this same relationship may not occur in pop, a curvilinear regression analysis was run on the difference scores for this genre. The beta weights showed that there was a significant inverted-U relationship between the difference scores and both general hit popularity (F (2, 58247) = 19.20, $p < .001$, energy beta = .06, energy-squared beta = -.05) and United Kingdom hit popularity (F (2, 58247) = 41.28, $p < .001$, energy beta = .05, energy-squared beta = -.02) indicating that the nature of this relationship is better characterized as curvilinear rather than monotonic.

- Table 2 about here -

Discussion

H1 stated that, given the predictions of Berlyne's theory, there should be an inverted-U relationship between measures of popularity and energy, indicating that moderately-arousing music is liked most. The data in Table 1 indicate that energy was indeed related to popularity, although the

nature of these relationships was not as predicted by Berlyne. Specifically, the relationships were U-shaped, such that moderately-energetic music was least popular and popularity increased towards the extremes of the energy dimension. It is noteworthy that the same pattern was found irrespective of the precise measure of popularity used. More simply, rather than favoring moderately arousing music, music sales appear to favor those pieces that would be relatively calming or exciting.

This raises the issue of how such findings, based on an entire corpus of music and music sales, might be reconciled with the considerable quantity of lab-based evidence using more limited samples of music and which indicates an inverted-U relationship between liking for this music and its arousal-evoking properties. While there have been failed attempts to identify this inverted-U relationship (reviewed above), we believe that it would be too strong a conclusion to simply give precedence to these and to the current, more comprehensive data set, and dismiss Berlyne's theory. Rather the current findings do not preclude the legitimacy of findings that the very extreme energy levels endemic to much of the music employed in earlier experimental work are disliked: extremely repetitive or apparently unordered, unstructured tones are undeniably unpopular relative to more 'conventional' music. However, in the context of the more restricted range of energy endemic to the music people actually buy, and which are represented in the current dataset, the relationship between energy and popularity appears U-shaped.

This argument is not so heretical as might seem on the basis of a narrow reading of the laboratory-based literature on Berlyne's theory: rather, several studies conducted over the past two decades have indicated that digitization and the increasing portability of music arising from this has led to people actively using music in contextualized everyday listening in order to achieve polarized arousal-based goals - such as using calming music to relax or arousing music to provide a psychological lift - consistent with the notion that there is a U-shaped relationship between liking for music and its arousal-evoking properties in data with greater ecological validity. For example, we noted earlier that North and Hargreaves (2000a) found that lab-based participants riding an exercise bike would select arousing music that would further polarize arousal and help them

perform better, rather than selecting a calming version of the same piece that would moderate their arousal; whereas people relaxing would select the calming version of that same piece that would further polarize arousal in order to support this goal, and eschew the more arousing version of the same piece that would moderate arousal. Krause's more recent studies of music playlists on portable devices (Krause & North, 2014, Krause, North, & Hewitt, 2014) similarly showed that respondents used music to achieve polarized rather than moderate arousal states, such as jogging while listening to arousing music or listening to calming music in order to relax during the commute home from work in the early evening.

Hypothesis 2 stated that there should be a negative linear relationship between popularity and difference scores. The data were consistent with this in the case of general hit appearance and United Kingdom hit appearance (see Table 2); however, the relationship was not found when popularity was defined specifically in terms of general hit popularity or United Kingdom hit popularity. The greater success of the hit appearance variables compared to hit popularity perhaps reflects the less-confected nature of the former, or may indicate that typicality is related positively to whether a piece will reach the charts (captured directly by the hit appearance variables), but not to how well it will perform once on the chart (captured only by the hit popularity variables).

Hypothesis 2a stated that the positive relationship between popularity and typicality may not be found in the case of pop music, indicating the greater commercial success of distinctive pieces. Consistent with H2a, curvilinear regression analysis indicated that the more accurate description of the relationship concerning pop music is that both highly innovative and highly derivative pieces do not enjoy the same popularity as do moderately typical pieces. In the case of pop, commercial success appears to be associated with music that has a degree of distinctiveness relative to other pieces that would allow it to gain attention in the crowded marketplace.

In summary, the present findings indicate that arousal- and typicality-based approaches can characterize the popularity of pieces across an entire commercial musical culture inclusive of music spanning several decades. The findings concerning the arousal-based approach were less consistent

with previous laboratory results, although the most probable explanation of this is the use here of only commercially-successful music, and it is possible to reconcile the present findings concerning this music with those of previous research that has deliberately employed music representing a very wide range of arousal. Results concerning the typicality-based approach were more consistent with those of previous laboratory research, at least in the case of hit appearance, if not hit popularity. It was also interesting that these same analyses indicated that commercial success in pop music nonetheless appears to require some degree of distinctiveness relative to the market. Similarly, even though consistent with theory, it is disappointing that the findings concerning general hit appearance and United Kingdom hit appearance indicate that greater financial reward is associated with more typical and derivative rather than innovative music (and note that separate tests not reported above show that this occurred even for arguably the two most high art genres considered, namely classical music and opera).

Before concluding we should also comment on the archival approach used in the present research. Although there are obvious advantages of this, there are also two important limitations. First, the effect sizes identified here were very small. The relative importance that should be attached to statistical significance in the context of small effect sizes has of course been the subject of considerable debate in recent years (e.g., Nickerson, 2000). At the risk of over-simplification, the core argument is that effect size and probability are simply different concepts, such that one cannot be taken as implying the other, particularly in the case of large samples of data such as that employed in the present research. Nonetheless, we believe that, despite the small effect size here, the statistical significance of the results is itself important and impressive. Although the use of a large sample increases the odds of identifying a statistically significant result, the crucial point is that the trends identified here were found within a commercially-complete musical culture: they exist in a manner that can be interpreted within existing theory. Moreover, none but the most ardent adherents to the theories would argue that the theories addressed here are complete explanations of the popularity of music. Rather it seems almost certain that the factors proposed within the theories

would be mediated and/or moderated by a number of factors. These would likely include music industry marketing techniques and strategies, the business demands of radio airplay, and changes in musical fashion. Similarly, the reliance here on pre-existing data sources and computerised analysis inevitably means that the means by which arousal and typicality were operations lied here surely represents only a limited component of the more general concepts as described in the theories themselves. The important point is that the theories do appear to explain at least some small portion of the population-wide popularity of a wide range of music, and it is encouraging that significant relationships were obtained at all. We might also note that, as with the great majority of research on musical taste, the present findings are limited to a particular culture: although it could be argued that the present corpus represents one musical culture very well, it does not speak at all to any other. Note also that the theories tested here nonetheless claim a basis in fundamental principles in human motivation, and so we would expect the present results to be replicable in other cultures.

References

Alcalde, V., Ricard, J., Bonet, A., Llopis, A., & Marcos, J. (2008). U.S. Patent No. 20080021851. Washington, DC: U.S. Patent and Trademark Office.

Alcalde, V., Ricard, J., Bonet, A., Llopis, A., & Marcos, J. (2010). U.S. Patent No. 20100250471. Washington, DC: U.S. Patent and Trademark Office.

Berlyne, D. E. (1971). *Aesthetics and psychobiology*. New York: Appleton-Century-Crofts.

Bornstein, R. F., & D'Agostino, P. R. (1994). The attribution and discounting of perceptual fluency: preliminary tests of a perceptual fluency/attributional model of the mere exposure effect. *Social Cognition*, *12*, 103–128.

Crozier, J. B. (1974). Verbal and exploratory responses to sound sequences varying in uncertainty level. in D. E. Berlyne (ed.), *Studies in the new experimental aesthetics* (pp. 27-90). New York: Wiley.

Fechner, G. T. (1876). *Vorschule der äesthetik*. Leipzig: Breitkopf and Hartel.

Hargreaves, D. J. (1986). *The developmental psychology of music*. Cambridge: Cambridge University Press.

Hekkert, P., & van Wieringen, P. C. W. (1990). Complexity and prototypicality as determinants of the appraisal of cubist paintings. *British Journal of Psychology*, *81*, 483-495.

Henrich, J. Heine, S.J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 61-83.

Krause, A. E., & North, A. C. (2014). Contextualized music listening: Playlists and the Mehrabian and Russell model. *Psychology of Well-Being: Theory Research and Practice*, *4*: 22.

Krause, A. E., North, A. C., & Hewitt, L. Y. (2014). The role of location in everyday experience of music. *Psychology of Popular Media Culture*, *10*, 3, np.

McMullen, P. T. (1974). Influence of number of different pitches and melodic redundancy on preference responses. *Journal of Research in Music Education*, *22*, 198-204.

McMullen, P. T., & Arnold, M. J. (1976). Preference and interest as a function of distributional redundancy in rhythmic sequences. *Journal of Research in Music Education*, *24*, 22-31.

Martindale, C. (1990). *The clockwork muse: the predictability of artistic change*. New York: Basic Books.

Martindale, C. (2007). Recent trends in the psychology of aesthetics, art, and creativity. *Empirical Studies of the Arts*, *25*, 121-141.

Martindale, C., & Moore, K. (1989). Relationship of musical preference to collative, ecological, and psychophysical variables. *Music Perception*, *6*, 431-455

Martindale, C., Moore, K., & Borkum, J. (1990). Aesthetic preference: anomalous findings for Berlyne's psychobiological theory. *American Journal of Psychology*, *103*, 53-80.

Martindale, C., Moore, K., & West, A. (1988). Relationship of preference judgements to typicality, novelty, and mere exposure. *Empirical Studies of the Arts*, *6*, 79-96.

Moore, K., & Martindale, C. (1983). Preference for shapes varying in color, color typicality, size, and complexity. *Paper presented at the International Conference on Psychology and the Arts, Cardiff.*

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241-301.

North, A. C., & Hargreaves, D. J. (1995). Subjective complexity, familiarity, and liking for popular music. *Psychomusicology*, *14*, 77-93.

North, A. C., & Hargreaves, D. J. (2000a). Musical preference during and after relaxation and exercise. *American Journal of Psychology*, *113, 43-67.*

North, A. C., & Hargreaves, D. J. (2000b). Collative variables versus prototypicality. *Empirical Studies of the Arts*, *18*, 13-17.

North, A. C. and Hargreaves, D. J. (2008). *The social and applied psychology of music*. Oxford: Oxford University Press.

North, A. C., Krause, A. E., Sheridan, L. P., & Ritchie, D. (2015). Energy and emotion in music: A computerized analysis of 143,353 pieces. *Manuscript submitted for publication*.

Russell, P. A. (1987). Effects of repetition on the familiarity and likeability of popular music recordings. *Psychology of Music*, *15*, 187-197.

Seamon, J. G., Brody, N., & Kauff, D. M. (1983). Affective discrimination of stimuli that are not recognized: Effects of shadowing, masking, and cerebral laterality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*, 544–555.

Simon, C. R., & Wohlwill, J. F. (1968). An experimental study of the role of expectation and variation in music. *Journal of Research in Music Education*, *16*, 227-238.

Simonton, D. K. (1980). Thematic fame, melodic originality, and musical zeitgeist: a biographical and transhistorical content analysis. *Journal of Personality and Social Psychology*, *38*, 972-983.

Vitz, P. C. (1966). Affect as a function of stimulus variation. *Journal of Experimental Psychology*, *71*, 74-79.

Whitfield, T. W. A. (1983). Predicting preference for familiar, everyday objects. An experimental confrontation between two theories of aesthetic behaviour. *Journal of Environmental Psychology*, *3*, 221-237.

Whitfield, T. W. A., & Slatter, P. E. (1979). The effects of categorization and prototypicality on aesthetic choice in a furniture selection task. *British Journal of Psychology*, *70*, 65-75.

Wintre, M. G., North, C., & Sugar, L. A. (2001). Psychologists' response to criticisms about research based on undergraduate participants: A developmental perspective. *Canadian Psychology*, *42*, 216–225.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, *9*, 1–27.

Zajonc, R. B. (1980). Feeling and thinking: preferences need no inferences. *American Psychologist*, *35*, 151-175.

Zajonc, R. B. (2001). Mere exposure: a gateway to the subliminal. *Current Directions in Psychological Science*, *10*(6), 224-228.

Table 1.
Quadratic Curvilinear Regression Results for the Analyses Testing Berlyne's Inverted-U Relationship

Tested dependent variable	r^2	*F*	Energy		Energy squared	
			Beta	*t*	Beta	*t*
General hit popularity	0.000	31.29***	-0.02	-1.30	0.04	3.12**
UK hit popularity	0.002	136.66***	-0.06	-5.19***	0.10	8.76***
General hit appearance	0.000	6.81**	-0.04	-3.40**	0.04	3.64***
UK hit appearance	0.000	26.74***	-0.08	-7.11***	0.07	6.50***

Note. Degrees of freedom = (2, 143350).
* $p < .05$, ** $p < .01$, *** $p < .001$

Table 2.
Correlation Coefficients Between the Total Difference Score and Measures of Popularity

Total difference score	General hit popularity	General hit appearance	UK hit popularity	UK hit appearance
Overall corpus (N = 143353)	.005	-.010***	.008**	-.011***

* $p < .05$, ** $p < .01$, *** $p < .001$