Parncutt, R., & Sattmann, S. (Eds.) (2018). *Proceedings of ICMPC15/ESCOM10.* Graz, Austria: Centre for Systematic Musicology, University of Graz.

Modeling Aural Skills Dictation

David John Baker,[1] Elizabeth Monzingo,[2] and Daniel Shanahan[3,4]

[1,2,3] *School of Music, Louisiana State University, United State of America*

[4]*School of Music, Ohio State University, United States of America*

[1]dbake29@lsu.edu, [2]emonzi1@lsu.edu, [3]daniel.shanahan@gmail.com

Abstract

Despite its abundance in curricula in music conservatory settings, research on topics pertaining to aural skills is currently limited at best. While anthologies of materials for sight singing and dictation exist, the ways in which people learn melodies are not well understood. This problem is difficult to tackle given the amount of factors that may contribute to the process, such as the complexity of the melody, the degree of exposure needed to commit a melody to long-term memory, and individual differences in cognitive ability that have been shown to contribute to an individual's performance on musical tasks. Fortunately, literature exists in related areas that serve to inform which parameters might contribute to an individual's performance in a melodic dictation setting. This paper presents findings from an experiment (N=39) modeling performance on melodic dictation tasks using both individual and musical features. Results suggest tools from computational musicology as well as individual difference measures need further exploration in order to assess the degree to which various features contribute to melodic dictation performance and inform pedagogical practices.

Introduction

Despite its near ubiquity in Conservatory and School of Music curricula, research surrounding topics concerning aural skills is not well understood. This is peculiar since almost any individual seeking to earn a degree in music usually must enrol in multiple aural skills classes which cover a wide array of topics from sight-singing melodies, to melodic and harmonic dictation– all of which are presumed to be fundamental to any musician's formal training. Skills acquired in these classes are meant to hone the musician's ear and enable them not only to think about music, but, to borrow Gary Karpinski's phrase, to "think in music" (Karpinski, 2000, p.4). The tacit assumption behind these tasks is that once one learns to think *in* music, these abilities should transfer to other aspects of the musician's playing in a deep and profound way. The skills that make up an individual's aural skills encompass many abilities, though are thought to be reflective of some sort of core skill. This is evident in early attempts to model performance in aural skills classes where C. S. Harrison, Asmus, and Serpe (1994) created a latent variable model to predict an individual's success in aural skills classes based on musical aptitude, musical experience, motivation, and academic ability. While their model was able to predict a large amount of variance (73%), modeling at this high, conceptual of a level does not provide any sort of specific insights into the mental processes that are required for completing aural skills related tasks. This trend can also be seen in more recent research that has explored the relationship between how well entrance exams at the university level are able to predict success later on in the degree program. Wolf and Kopiez (2014) noted a multiple confounds in their study attempting to asses ability level in university musicians such as inflated grading, which led to ceiling effects, as well as a broad lack of consistency in how schools are assessing success within their students. But even if the results at the larger level were to be clearer, again this says nothing about the processes that contribute to tasks like melodic dictation. Rather than taking a bird's eye view of the subject, this paper will primarily focus on factors that might contribute to an individual's ability dictate a melody.

Melodic dictation is one of the central activities in an aural skills class. The activity normally consists of the instructor of the class playing a monophonic melody a limited number of times and the students must use both their ears, as well as their understanding of Western Music theory and notation, in order to transcribe the melody without any sort of external reference. No definitive method is taught across universities, but many schools of thought exist on the topic and a wealth of resources and materials have been suggested that might help students better complete these tasks (Berkowitz, Frontier, & Kraft, 1960; Cleland & Dobrea-Grindahl, 2013; Karpinski, 2007; Ottman, 1996). The lack of consistency could be attributed to the fact that there are so many processes at play during this process. Prior to listening, the student needs to have an understanding of Western music notation at least to the degree of understanding of the melody being played. This understanding needs to be readily accessible, since as new musical information is heard, it is the student's responsibility to, in that moment, encode the melody into either hold a chunk of the melody in short term memory or pattern match to long term memory so that they can identify what they are hearing and transcribe it moments later into Western notation. So no matter what, performing some sort of aural skills task requires both long term memory and knowledge for comprehension, as well as the ability to actively manipulate differing degrees of complex musical information in real time while concurrently writing it down. Given this complexity of the task, as well as the difficulty in quantifying attributes of melodies, it is then not surprising that scant research exists on describing these tasks.

Fortunately, a fair amount of research exists in related literature which can generate theories and hypotheses explaining how individuals dictate melodies. Beginning first with factors that are less malleable from person to person would be individual differences in cognitive ability. While dictating melodies is something that is learned, a growing body of literature suggests that other factors can explain unique amounts of variance in performance via differences in cognitive ability. For example, Meinz and Hambrick (2010) found that measures of working memory capacity (WMC)

were able to explain variance in an individual's ability to sight read above and beyond that of sight reading experience and musical training. Colley, Keller, and Halpern (2017) recently suggested an individual's WMC also could help explain differences beyond musical training in tasks related to tasks of tapping along to expressive timing in music. These issues become more confounded when considering other recent work by Swaminathan, Schellenberg, and Khalil (2017) that suggests factors such as musical aptitude, when considered in the modeling process, can better explain individual differences in intelligence between musicians and non-musicians implying that within the musical population. They claim there is a selection bias that "smarter" people tend to gravitate towards studying music, which may explain some of the differences in memory thought to be caused by music study (Talamini, Altoè, Carretti, & Grassi, 2017). Knowing that these cognitive factors can play a role warrants attention from future researchers on controlling for variables that might that might contribute to this process but are not directly intuitive and have not been considered in much of the past research. This is especially important given recent critique of models that purport to measure cognitive ability but are not grounded in an explanatory theoretical model (Kovacs & Conway, 2016).

The ability to understand how individuals encode melodies is at the heart of much of the music perception literature. Largely stemming from the work of Bregman (1994), Deutsch and Feroe (1981), and Dowling (1978; 1971) work on memory for melodies has begun to lay the foundation for how people learn melodies. Initial work by Dowling suggested that both key and contour information play a central role in the perception and memory of novel melodies. Interestingly enough, memory for melodies tends to be much worse than memory for other stimuli such as pictures or faces noting that the average area under the ROC curve tends to be at about .7 in many of the studies they reviewed, with .5 meaning chance and 1 being a perfect performance (Halpern and Bartlett,2010). Halpern and Bartlett also note that much of the literature on memory for melodies primarily used same difference experimental paradigms to investigate individual's melodic perception ability similar to the paradigm used in Halpern and Müllensiefen (2008).

Not nearly as much is known about how an individual learns melodies, especially in dictation setting. The last, and possibly most obvious, variable that would contribute to an individual's ability to learn and dictate a melody would be the amount of exposure to the melody and the complexity of the melody itself. There is not much research on the first of these two points, other than an approximation of how many times the melody should be played in a dictation setting according to (Karpinski, 2007, p.100) that accounts for chunking as well as the idea that more exposure would lead to more complete encoding.

Recently tools have been developed in the field of computational musicology to help with operationalizing how complex melodies are. Both simple and more complex features have been used to model performance in behavioral tasks. For example Eerola, Himberg, Toiviainen, and Louhivuori (2006) found that note density, though not consciously aware to the participants, predicted judgments of human similarity between melodies not familiar to the participants. Note density would be an ideal candidate to investigate as it is both easily measured and the amount of information that can be currently held in memory as measured by bits of information has a long history in cognitive psychology (Cowan, 2015; Miller, 1956)

In terms of more complex features, much of the work largely stems from the work of Müllensiefen and his development of the FANTASTIC Toolbox (2009), a few papers have claimed to be able to predict various behavioral outcomes based on the structural characteristics of melodies. For example, Kopiez and Müllensiefen (2011) claimed to have been able to predict how well songs from The Beatles' album *Revolver* did on popularity charts based on structural characteristic of the melodies using a data driven approach. Expanding on an earlier study, Müllensiefen and Halpern (2014) found that the degree of distinctiveness of a melody when compared to its parent corpus could be used in order to predict how participants in an old/new memory paradigm were able to recognize melodies.

These abstracted features also have been used in various corpus studies (Frieler, Jakubowski, & Müllensiefen, 2015; Jakubowski, Finkel, Stewart, & Müllensiefen, 2017; Janssen, Burgoyne, & Honing, 2017; Rainsford, Palmer and Paine 2017). that again use a machine learning approach in order to explain which of the 38 features that FANTASTIC calculates can predict real-world behavior. In addition to looking at individual features, or sets of features, as predictors, recent work by P. Harrison, Musil, and Müllensiefen (2016), Baker and Müllensiefen (2017) and the aforementioned Müllensiefen and Halpern (2014) study have used data reduction techniques, namely principal component analysis, to take measures that were successful in predicting behavioral outcomes and boil them down into a single measure of complexity that has had predictive power in modeling experimental performance. While helpful and somewhat explanatory, the problem with many of these approaches is that they take a post-hoc data driven approach with the assumption that listeners are even able to abstract and perceive these features. Doing this does not allow for any sort of controlled approach and without experimentally manipulating the parameters, which is then further confounded when using some sort of data reduction technique. This is understandable seeing as it is very difficult to manipulate certain qualities of a melody without disturbing other features. For example, if you wanted to decrease the "tonalness" of a melody by adding in a few more chromatic pitches, you inevitably will increase other measures of pitch and interval entropy. In order to truly understand if these features are driving changes in behaviour, each needs to be altered in some sort of controlled and systematic way while simultaneously considering differences in training and cognitive ability.

Aims

This paper presents findings from two experiments modeling performance on melodic dictation tasks using both individual and musical features. A pilot study was run (N=11) was used in order to assess musical confounds that might be present in modeling melodic dictation. Results of that pilot study are not

reported here. Based on the results of this pilot data, a follow up experiment was conducted to better investigate the features in question.

The study sought to answer two main hypotheses:

- H1: Are all experimental melodies used equally difficult to dictate?
- H2: To what extent do the musical features of Note Density and Tonalness play a role in difficulty of dictation?
- H3: Do individual factors at the cognitive level play a role in the melodic dictation process above and beyond musical factors?

Methods

Participants Forty-three students enrolled at Louisiana State University School of Music completed the study. The inclusion criteria in the analysis included reporting no hearing loss, not actively taking medication that would alter cognitive performance, and individuals whose performance on any task performed greater than three standard deviations from the mean score of that task. Using these criteria two participants were dropped for not completing the entire experiment. Thus, 41 participants met the criteria for inclusion. The eligible participants were between the ages of 17 and 26 (M = 19.81, SD = 1.93; 15 women). Participants volunteered, received course credit, or were paid $10.

Materials Four melodies for the dictation were selected from a corpus of n=115 melodies derived from the *A New Approach to Sight Singing* aural skills textbook by Berkowitz et. al (2005). Melodies were chosen based on their musical features as extracted via the FANTASTIC Toolbox (Müllensiefen, 2009). After abstracting the full set of features of the melodies, possible melodies were first narrowed down by limiting the corpus to melodies lasting between 9 and 12 seconds and then indexed to select four melodies were chosen that as part of a 2x2 repeated measures design including a high and low tonalness and note density condition. Melodies, as well as a table of their abstracted features can be seen in Table 1 and Figures 1—4. Melodies and other sounds used were encoded using MuseScore 2 using the standard piano timbre and all set to a tempo of quarter = 120 beats per minute and adjusted accordingly based on time signature to ensure they all sounded the same absolute time duration. The experiment was then coded in jsPsych (de Leeuw, 2015) andaccesed through a browser offline with high quality headphones.

Table 1. Features of Melodies as Computed by FANTASTIC.

Melody	Note Density (ND)	Tonalness	Label
9	1.75	.71	Low ND, Low Tonal
34	1.66	.94	Low ND, High Tonal
95	3.91	.76	High ND, Low Tonal
112	3.73	.98	High ND, High Tonal

Figure 1: Melody 9

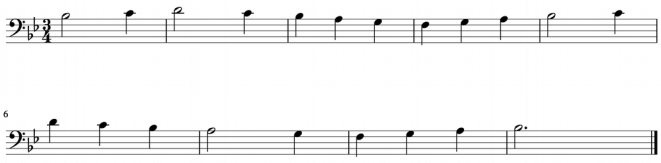


Figure 2: Melody 34

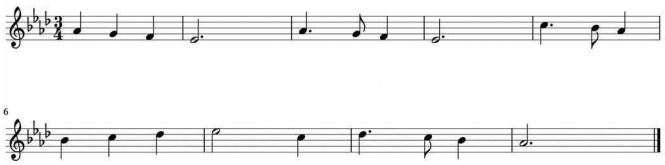


Figure 3: Melody 95

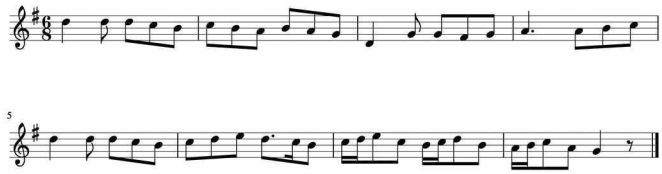
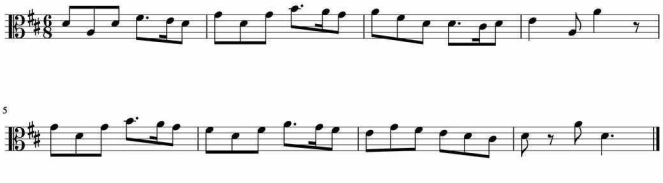


Figure 4: Melody 112



Procedure Upon arriving at the lab, participants sat down in a lab at their own personal computer. Multiple individuals were tested simultaneously although individually. Each participant was given a test packet which contained all information needed for the experiment. After obtaining written consent participants navigated through a series of instructions explaining the nature of the experiment and given an opportunity to adjust the volume to a comfortable level. The first portion of the experiment that participants completed was the melodic dictation. In order to alleviate any anxiety in performance, participants were explicitly told that "unlike dictations performed in class, they were not expected to get perfect scores on their dictations". Each melody was played five times with 20 seconds between hearings and 120 seconds after the last hearing. After the dictation portion of the experiment, participants completed a small survey on their Aural Skills background, as well as the Bucknell Auditory Imagery Scale C (Halpern, 2015). After completing the Aural Skills portion of the experiment participants completed one block of two different tests of working memory capacity (Unsworth et al., 2005) and Raven's Advanced Progressive Matrices and a Number Series task as two tests of general fluid intelligence (Gf) (Raven et al., 1998; Thurstone, 1938) resulting in four total scores. After completing the cognitive battery, participants finished the experiment by compiling the self-report version of the Goldsmiths Musical Sophistication

Index (Müllensifen et. al, 2014), the Short Test of Musical Preferences (Rentfrow & Gosling, 2003), as well as questions pertaining to the participants SES, and any other information we needed to control for (Hearing Loss, Medication). Exact materials for the experiment can be found at https://github.com/davidjohnbaker1/modelingMelodicDictation .

Scoring Melodies were scored by counting the amount of notes in the melody and multiplying that number by two. Half the points were attributed to rhythmic accuracy and the other half to pitch accuracy. Points were not deducted for notating the melody in the incorrect octave. Points for pitch could only be given if the participant correctly notated the rhythm. For example, in melody 34 there were 40 points possible (20 notes * 2). If a participant were to have put a quarter note on the second beat of the third measure, and have everything else correct, they would have scored a 19/20. Only if the correct rhythms of the measures were accurate could pitch points be awarded. In cases where there were more serious errors, for example if the second half of the second bar was not notated, points would have been deducted in both the pitch and rhythm sub-scores. Both the first and second author scored all melodies independently and then cross referenced for inter-rater reliability. Using a single score intraclass correlation coefficient calculation $\kappa = .96$ which suggests a high degree of inter-rater reliability (McHugh, 2012).

Results

Data Screening

Before conducting any analyses data was screened for quality. List wise deletion was used to remove any participants that did not have all variables used in modeling. This process resulted in removing four participants: two did not complete any of the survey materials and two did not have any measures of working memory capacity due to computer error. After list-wise deletion, thirty-nine participants remained.

Effects of Melodic Features

In order to investigate H_1, that melodies would differ in their degree of difficulty based on melodic features, we ran a repeated measures ANOVA using the *ez* package in R (Lawrence, 2016). Relevant statistics from the model can be seen in Table 2.

Table 2: Repeated Measures ANOVA

Predictor	df_{Num}	df_{Den}	SS_{Num}	SS_{Num}	F	p	η_g
(Intercept)	1	38	19.18	3.85	189.21	.000	.73
Tonalness	1	38	0.10	0.83	4.37	.043	.01
NoteDensity	1	38	5.99	1.47	154.90	.000	.46
Tonalness x NoteDensity	1	38	0.15	0.82	6.88	.012	.02

Subsequent models exploring possible exploratory covariance relationships using random slope models that used measures of working memory capacity, general fluid intelligence, and measures of musical training, none of which emerged as significant.

Differences between melodies can be see below in Figure 5.

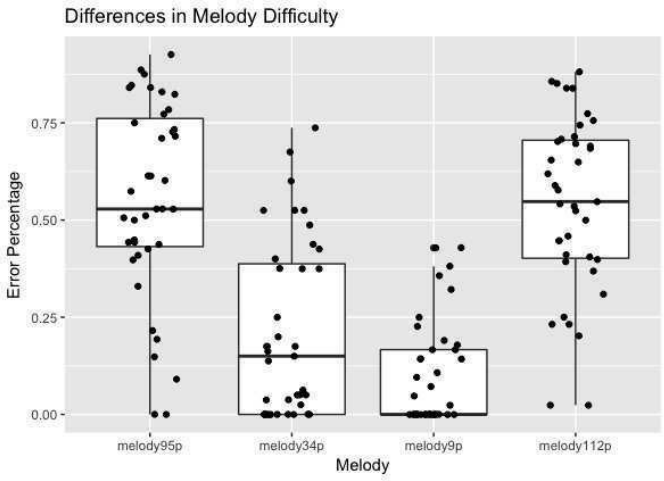


Figure 5: Boxplot of Differences between Melodies

Discussion

Here, we have investigated the extent to which both individual differences and abstracted musical features could be used to model results in melodic dictations. In order to examine H1, we ran a repeated measures ANOVA in order discern any differences in melody difficulty. As noted in Table 2, both a significant main effect of Tonalness and Note Density was found, as well as a small interaction between the two variables suggesting evidence supporting rejecting H2's null hypothesis. The interaction emerged from differences in melody means in the low density conditions with the melody with higher tonalness actually scoring higher in terms of number of errors.

While we expected to find an interaction, this condition (Melody 34) was hypothesized to be the easiest of the four conditions. With Melody 9 there was a clear floor effect, which was also to be expected as when we chose the melodies, we had no previous experimental data explicitly looking at melodic dictation to rely on. For future experiments, we will use abstracted features from Melody 9 as a baseline. The main effect of note density was expected and exhibited a large effect size. ($\eta_g = .46$). While it would be tempting to attribute this finding exactly to the Note Density feature extracted by FANTASTIC, the high and low density conditions could also be operationalized as having compound versus simple meter. Given the large effect of note density, we plan on taking more careful steps in the selection of our next melodies in order to control for any effects of meter and keep the effects limited to one meter if at all possible.

Somewhat surprisingly, the analysis incorporating the cognitive measures of covariance did not yield any significant results. While other researchers have noted the importance of baseline cognitive ability (Schellenberg & Weiss, 2013), the task specificity of doing melodic dictation as we designed the experiment might not be well suited to capture the variability needed for any effects. Hence, this paper would not be able to reject H_3's null hypothesis. Considering that other researchers have founding constructs like working memory capacity and general fluid intelligence to be important factors of tasks of musical perception, a more refined design might be considered in the future to find any sort of effects.

Taken as a whole, these findings suggest that aural skills pedagogues should consider exploring the extent to which computationally extracted features can guide the difficulty expected of melodic dictation exercises.

Conclusion

This paper demonstrates that abstracted musical features such as tonalness and note density can play a role in predicting how well students do in tasks of melodic dictation. While the experiment failed to yield any significant differences in cognitive ability predicting success at the task, our future research plans to continue incorporate measures that others have deemed important. We next plan to replicate this experiment's design with different melodies that use similar features.

Acknowledgements. The authors would like to thank Adam Rosado, Brian Ritter, and Katherine Vukovics for helping run participants on this study. Additionally, the authors would like thank Dr. Emily Elliott for providing hardware and software required to run the tests of working memory capacity and general fluid intelligence.

References

Baker, D. J., & Müllensiefen, D. (2017). Perception of leitmotives in Richard Wagner's der Ring des Nibelungen. Frontiers in Psychology, 8.

Berkowitz, S., Frontrier, G., & Kraft, L. (1960). A new approach to sight singing. WW Norton.

Bregman, A. S. (1994). Auditory scene analysis: The perceptual organization of sound. MIT press.

Cleland, K. D., & Dobrea-Grindahl, M. (2013). Developing musicianship through aural skills: A holistic approach to sight singing and ear training. Routledge.

Colley, I. D., Keller, P. E., & Halpern, A. R. (2017). Working memory and auditory imagery predict sensorimotor synchronization with expressively timed music. The Quarterly Journal of Experimental Psychology(just-accepted), 1–49.

Cowan, N. (2015). George miller's magical number of immediate memory in retrospect: Observations on the faltering progression of science. Psychological review, 122 (3), 536.

Deutsch, D., & Feroe, J. (1981). The internal representation of pitch sequences in tonal music. Psychological review, 88 (6), 503.

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior research methods*, *47*(1), 1-12.

Dowling, W. J. (1978). Scale and contour: Two components of a theory of memory for melodies. Psychological review, 85 (4), 341.

Dowling, W. J., & Fujitani, D. S. (1971). Contour, interval, and pitch recognition in memory for melodies. The Journal of the Acoustical Society of America, 49 (2B), 524–531.

Eerola, T., Himberg, T., Toiviainen, P., & Louhivuori, J. (2006). Perceived complexity of western and african folk melodies by western and african listeners. Psychology of Music, 34 (3), 337–371.

Frieler, K., Jakubowski, K., & Müllensiefen, D. (2015). Is it the song and not the singer? Hit song prediction using structural features of melodies. Yearbook of Music Psychology (Jahrbuch Musikpsychologie).

Halpern, A. R. (2015). Differences in auditory imagery self-report predict neural and behavioral outcomes. *Psychomusicology: Music, Mind, and Brain*, *25*(1), 37.

Halpern, A. R., & Bartlett, J. C. (2010). Memory for melodies. In Music perception (pp. 233–258). Springer.

Halpern, A. R., & Müllensiefen, D. (2008). Effects of timbre and tempo change on memory for music. The Quarterly Journal of Experimental Psychology, 61 (9), 1371–1384.

Harrison, C. S., Asmus, E. P., & Serpe, R. T. (1994). Effects of musical aptitude, academic ability, music experience, and motivation on aural skills. Journal of Research in Music Education, 42 (2), 131–144.

Harrison, P., Musil, J. J., & Müllensiefen, D. (2016). Modelling melodic discrimination tests: descriptive and explanatory approaches. Journal of New Music Research, 45 (3), 265–280.

Huron, D. B. (1994). The humdrum toolkit: Reference manual. Center for Computer Assisted Research in the Humanities.

Jakubowski, K., Finkel, S., Stewart, L., & Müllensiefen, D. (2017). Dissecting an earworm: Melodic features and song popularity predict involuntary musical imagery. Psychology of Aesthetics, Creativity, and the Arts, 11 (2), 122.

Janssen, B., Burgoyne, J. A., & Honing, H. (2017). Predicting variation of folk songs: A corpus analysis study on the memorability of melodies. Frontiers in Psychology, 8 .

Karpinski, G. S. (2000). Aural skills acquisition: The development of listening, reading, and performing skills in college-level musicians. Oxford University Press on Demand.

Kopiez, R., & Müllensiefen, D. (2011). Auf der suche nach den popularitätsfaktoren in den song-melodien des beatles-albums revolver eine computergestützte feature-analyse. na.

Kovacs, K., & Conway, A. R. (2016). Process overlap theory: a unified account of the general factor of intelligence. Psychological Inquiry, 27 (3), 151–177.

Michael A. Lawrence (2016). ez: Easy Analysis and Visualization of Factorial Experiments. R package version 4.4-0. https://CRAN.R-project.org/package=ez

McHugh ML. Interrater reliability: the kappa statistic. *Biochemia Medica*. 2012;22(3):276-282.

Meinz, E. J., & Hambrick, D. Z. (2010). Deliberate practice is necessary but not sufficient to explain individual differences in piano sight-reading skill: The role of working memory capacity. Psychological science, 21 (7), 914–919.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological review, 63 (2), 81.

Müllensiefen, D. (2009). Fantastic: Feature analysis technology accessing statistics (in a corpus): Technical report v1. Goldsmiths University of London.

Müllensiefen, D., & Frieler, K. (2006). The simile algorithms documentation 0.3. White Paper.

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

Müllensiefen, D., & Halpern, A. R. (2014). The role of features and context in recognition of novel melodies. Music Perception: An Interdisciplinary Journal, 31 (5), 418–435.

Müllensiefen, D., Wiggins, G., Lewis, D., et al. (2008). High-level feature descriptors and corpus-based musicology: Techniques for modelling music cognition. na.

Ottman, R. W. (1996). Music for sight singing. Prentice Hall.

Rainsford, M., Palmer, M., & Paine, G. (2017). The musos (music software system) toolkit: A computer-based, open source application for testing memory for melodies. Behavior Research Methods, 1–19.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales.* Oxford, England: Oxford Psychologists Press.

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.

Schellenberg, E. G., & Weiss, M. W. (2013). Music and cognitive abilities. In *The Psychology of Music (Third Edition)* 499-550

Swaminathan, S., Schellenberg, E. G., & Khalil, S. (2017). Revisiting the association between music lessons and intelligence: Training effects or music aptitude? Intelligence.

Talamini, F., Altoè, G., Carretti, B., & Grassi, M. (2017). Musicians have better memory than nonmusicians: A meta-analysis. PloS one, 12 (10), e0186773.

Team, R. C., et al. (2013). R foundation for statistical computing. Vienna, Austria, 3 (0).

Thurstone, L. L. (1938). Primary mental abilities. Chicago, IL: University of Chicago Press.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. Behavior research methods, 37 (3), 498–505.

Wolf, A., & Kopiez, R. (2014). Do grades reflect the development of excellence in music students? the prognostic validity of entrance exams at universities of music. Musicae Scientiae, 18 (2), 232–248.