Recognition of leitmotives in Richard Wagner's music: An item response theory approach

Daniel Müllensiefen[1], David Baker[1], Christophe Rhodes[1], Tim Crawford[1], and Laurence Dreyfus[2]

[1] Goldsmiths, University of London, New Cross, SE14 6NW, United Kingdom
{d.mullensiefen,d.baker,c.rhodes,t.crawford}@gold.ac.uk
[2] University of Oxford, Wellington Square, Oxford OX1 2JD, United Kingdom
laurence.dreyfus@magd.ox.ac.uk

Abstract. In this study we aim to understand listeners' real-time processing of musical leitmotives. We probe participants' memory for different leitmotives contained in a 10-minute passage from the opera *Siegfried* by Richard Wagner, and use item response theory to estimate parameters for item difficulty and for participants' individual recognition ability, as well as to construct novel measurement instruments from questionnaire-based tests. We investigate the relationship between model parameters and objective factors, finding that prior Wagner expertise and musical training were significant predictors of leitmotive recognition ability, while item difficulty is explained by chroma distance and perceived emotional content of the leitmotives.

1 Introduction

1.1 Psychology of Leitmotives

The leitmotives in Richard Wagner's *Der Ring des Nibelungen* serve a range of compositional and psychological functions, including the introduction of musical structure and mnemonic devices for the listener. These leitmotives are short musical ideas that are representative of concepts in the dramatic narrative, and differ greatly in their construction, salient aspects (*e.g.* rhythmic, melodic, harmonic), and their usage in particular scenes and contexts. While the topic of leitmotives in Richard Wagner's music has been discussed extensively in the traditional musicological literature (Dalhaus, 1979), little work has been done on the perception and psychology of real-time processing of these musical ideas. In this study, we perform a psychological experiment to attempt to understand how individuals are able to recall leitmotives, investigating both musical- and listener-based parameters. Using an item response theory (IRT) approach, we estimate difficulty parameters of the leitmotives

(items) themselves, as well as parameters characterizing participants' individual recognition ability.

A small number of prior studies have empirically investigated the perception of leitmotives. Initial research on the leitmotives used a learning paradigm to explore how listeners with various musical backgrounds would encode and subsequently recognise various leitmotives in real time, using an excerpt from *Das Rheingold*, finding (Deliège, 1992) that musicians were able to encode musical material much more rapidly than non-musicians, and that each of the leitmotives presented different levels of difficulty in their recognition. This research was expanded upon by introducing additional visual stimuli, as well as considering listener parameters beyond musical training, finding (Albrecht, 2012) that visual stimuli did not help leitmotive recognition, but that the non-musical parameter of Wagner expertise did predict an individual's recognition ability. This study explores the difficulty of encoding the leitmotives and the contributions of extra-musical factors to an individual's recognition rate.

1.2 Experimental Design and Procedure

The experiment used a within-subjects design. Participants were asked to listen actively to the same ten-minute passage from Richard Wagner's opera *Siegfried* used by Albrecht (2012). This passage was chosen for its narrative qualities and high leitmotive density. The participants were told in advance that they would perform a memory recall task following the listening phase, in which they would have to indicate explicitly whether or not they recall hearing musical material from the passage, and to rate the subjectively perceived emotional qualities of the musical material, such as the level of emotional arousal and valence expressed. After the listening phase, participants were played a list of 20 excerpts, each containing a leitmotive. Ten of these leitmotives had occurred in the passage that they had heard before; the other ten were taken from a passage from the same performers' recording of Richard Wagner's *Götterdämmerung*. For each item participants had to indicate: whether they had heard the leitmotive in the 10-minute listening phase or not; how confident they were in their decision; and also how emotionally arousing they perceived the leitmotive together with an emotional judgement (happy-sad) both on 7-point scales.

After completing this memory recognition task, participants filled out questionnaires assessing factors that we believed might contribute to an individual's leitmotive recognition ability: musical training, measured using the Musical Training subscale of the Goldsmiths Music Sophistication Index self-report questionnaire (Müllensiefen *et al.*, 2014); affinity for the music of Richard Wagner and objective Wagner knowledge, measured with two novel questionnaire instruments that were constructed via Factor analysis and Rasch modelling (see Sections 2.2 and 2.3 below); and German speaking ability, on a 7-point agreement Likert scale.

1.3 Advantages of item response approaches in psychological research

Item response theory (Rasch, 1960; Birnbaum, 1968; Lord, 1980) was developed to assess individuals on attributes that are not directly observable ("latent" traits, such as aspects of intelligence or personality) using data from the individuals' performance on a suitable test. Among the most commonly cited advantages of IRT and latent trait models are: their foundation in well-established statistical theory (maximum-likelihood modelling); and their ability to quantify uncertainty via confidence intervals. In addition, Rasch models are a special class of IRT models which possess the property that item and person scores can be considered indepedent from the particular sample used.

Most concepts in cognitive psychology that are used to describe mental processes (such as memory capacity or attention span) are unobservable, yet item response approaches are still relatively rare within cognitive or experimental psychology. Borsboom (2006) discusses a number of reasons for the slow uptake of IRT models in most areas of psychology and also encourages its wider application. The current study represents a suitable scenario for IRT, where experimental data is generated by individuals taking a newly-designed test, and where the two main research questions investigate a) person-based factors explaining the individual's ability to perform on the test and b) per-item factors contributing to item difficulty. We are asking what characterizes listeners who perform better at encoding leitmotives in a realistic listening situation, and what musical or acoustic factors contribute to the recognizability of individual leitmotives. Compared to traditional analysis approaches in cognitive psychology, the IRT approach enables us to estimate participant ability and item difficulty within the same model and to quantify the uncertainty about both kinds of parameters through confidence intervals.

2 Data Analysis

2.1 Variables measuring participant background

As described in Section 1.2, we collected four person-specific pieces of information: musical expertise, German speaking ability, Wagner affinity and Wagner knowledge. The experiment used a convenience sample ($N = 100$), with additional recruiting effort made to recruit participants with either familiarity or fondness of the music of Richard Wagner from across the greater London area, though more ($N = 14$) individual's data was used in a pilot experiment and their survey and quiz response were retained for the final Rasch modelling ($N = 114$). The experimental ($N = 100$) sample was made up of 55 females (55%) and 45 males (45%) with a mean age of 28.7 (s.d. = 11.82). It is worth noting that the following analyses proceed in a step-wise fashion, where we first fit IRT and factor models to the data of the several tests and

questionnaire separately and aim to establish sound measurement models for each of these novel tests. Only subsequently we combine data in a structural equation model and a regression model. This step-wise analysis procedure allows us to check model assumptions at each stage and, where necessary, to apply adjustements to individual models (*e.g.* by excluding items that violate assumptions). However, the construction of the measurement models and the modelling of the structural relations between the factors of interest were carried out independently to avoid modelling bias.

2.2 Factor Analysis of Wagner Affinity Survey

To model individuals' affinity with Wagner, we do not need to assess item (question) difficulty, and so we can use factor analytic techniques, rather than having to apply a graded response model (Samejina, 1969) to the Likert-scale data of the survey. We conducted minimal residual factor analysis on the 14 items of the affinity questionnaire using the R psych package (Revelle, 2014). Parallel analysis (Horn, 1965) as well as Velicer's Simple Structure (Revelle and Rocklin, 1979) and the Minimum Average Partial correlation (Velicer, 1976) criterion were employed to decide on the number of factors, giving ambiguous results (suggesting either 1 or 2 factors). We inspected the items for their respective factor loadings on a 1-dimensional solution, finding that only one had a factor loading of less than 0.6 (with a loading of 0.482). After the removal of this item, "How often do you perform the music of Richard Wagner?", we reran the minimum residual factor analysis, and all the diagnostics suggested 1-dimensional factor solution. Cronbach's α for this solution was 0.97, indicating a high internal reliability of this new Wagner affinity scale in terms of classical test theory.

2.3 Rasch modelling of the Wagner Quiz

We designed the Wagner knowledge quiz to measure a postulated latent trait of "Wagnerism", the extent to which an individual has developed knowledge of the life and music of Richard Wagner both in terms of musicological knowledge as well as a detailed understanding of the narrative and music of his operas. The quiz had 14 multiple choice items, each with four response options, and each item was scored as either correct (1) or incorrect (0).

Because of the limited sample size ($N = 114$) we fit a Rasch model, a comparatively simple member of the family of IRT models (de Ayala, 2009), requiring relatively few parameters to be estimated. The initial Rasch model was fitted using the conditional maximum likelihood criterion as implemented in eRm package in R (Mair and Hatzinger, 2007) which assumes equal item difficulty as well as equal discrimination across the participant subgroups. However, Pononcy's non-parametric T_{10} (with median split sampling 1000 matrices) for subgroup invariance as well as the T_{pbis} test for equal item

discrimination both indicated that the assumptions were not met. Applying a stepwise elimination procedure based on individual item fit removed 6 items and resulted in a new Rasch model containing 8 items. This resulting model passed both the T_{10} and T_{pbis} tests as well a non-parametric version of the Martin-Loef (Glas and Verhelst, 1995) test for unidimensionality indicating that the main assumptions for Rasch models were met for the final 8-item model. The item difficulty parameters of the final version of the Wagner knowledge test showed a good range, from 1.04 for the item "When did Wagner die?" to -1.28 for the item "What opera is considered to be among the romantic operas that paved the way for Wagner's music dramas?".

2.4 Listening Response Analysis

The memory test contained 20 items, and participants responded with either a "yes" or "no" depending on whether they recognized the leitmotive from the 10-minute listening passage or not. Each response was scored using a binary coding as either correct (1) or incorrect (0). These binary responses were then analyzed by fitting using a Rasch Model for the same reasons as in Section 2.3. Applying the non-paremetric T_{pbis} test as implemented in the R package eRm (Mair and Hatzinger, 2007) to the model indicated an equal item discrimination, but the T_{10} test suggested that it was missing subgroup invariance. A graphical model check also indicated that several items differed in difficulty across the high and low performing subgroups of subjects. However, the result of the non-parametric Martin-Loef test suggested that the memory test would meet the criterion of unidimensionality.

The failure of the Rasch model to meet the criterion of subgroup invariance leaves several options. First, we explored fitting a two-parameter model with an additional guessing parameter per item (but equal discrimination) to accommodate for the possibility that participants were guessing on individual items, using the marginal maximum likelihood approach provided by the R package ltm (Rizopoulos, 2006). However, the two-parameter solution appeared to be degenerate with several difficulty parameters outside the normal range. Second, we considered excluding items from the test until model assumptions are met, as was done for the Wagner knowledge test, or alternatively modelling items with a multi-dimensional IRT model. But because the leitmotive items themselves are the objects of interest in one of the subsequent analysis stages, excluding several items from the small initial pool of only 20 would leave too few to generate interesting results in terms of the memorability of different types of leitmotives. Therefore, we opted to accept the existing model, acknowledging that one of the model assumptions is not met. This means that there is some uncertainty about the item difficulty parameters. However, as Lord (1980, p. 190) points out, the use of a Rasch model might still be justified when the sample size is small, even if assumptions do not hold. In this case, estimators derived from the Rasch model might not be optimal, but might still be more accurate than estimators derived from more flexible

IRT models (*e.g.* the 3-parameter model).

2.5 Modelling memory for leitmotives with a structural equation model

We specified a structural equation model (SEM) to determine the contributions of person-specific variables to explain the memory performance in the leitmotive recognition test. The person parameters from the Rasch model for the memory test (see Section 2.4) served as the target variable. As predictor variables we specified the musical training and German speaking scores, and a latent Wagner expertise variable, hypothesised to influence Wagner knowledge and affinity scores (which we treated as observed variables in the context of this SEM). We also specified covariances between Gold-MSI scores and Wagner knowledge as well as Wagner affinity scores. This initial model was entirely hypothesis driven and was fit using the R package lavaan using maximum likelihood with robust standard errors (Yves, 2012).

The initial model already showed an almost acceptable fit as suggested by the fit indices derived from the robust estimator (Comparative Fit Index = .94; Tucker-Lewis Index = .8, RMSEA = .16, SRMR = .07). We inspected the model parameters and removed one non-significant regression path (from German speaking ability to memory scores) and one non-significant covariance (between Musical Training and Wagner knowledge). We refit the model, resulting in a model with only significant path coefficients and showing almost perfect fit indices (CFI = 1; TLI = 1, RMSEA < .001, SRMR = .01). The model, depicted in Figure 1, shows that Wagner expertise and musical training both positively influence the participants ability to recognise leitmotives in the listening test; Wagner expertise is about twice as influential as musical training.

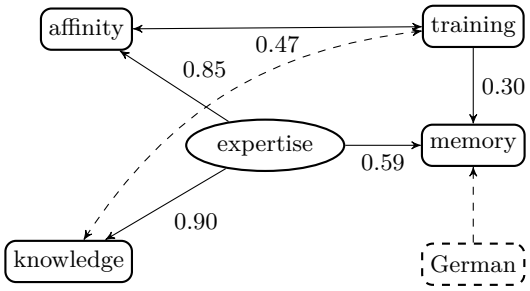

Fig. 1. Structural Equation Model for memory of leitmotives, incorporating Wagner knowledge and affinity, their combination into Wagner expertise, and the effect of that and generic musical training on score in the memory test. The dashed lines and boxes indicate non-significant relations removed from the final model, which contains only significant influences.

2.6 Modelling leitmotive difficulty

Previous evidence in the literature (Müllensiefen and Halpern, 2014) suggests that different musical features are responsible for the correct recognition of previously heard melodies ('hits') and the correct identification as novel of melodies that have not been previously heard ('correct rejections'). Therefore, we split the set of 20 leitmotives into 'old' motives (that had been heard previously in the experiment) and 'novel' motives (that did not occur in the passage), and ran two separate linear regression analyses with the item difficulty scores from the Rasch model as dependent variables. In both models the predictor variables were the mean of the participants' arousal and valence ratings carried out at the recognition phase as well as an acoustical distance measure based on chroma feature extraction (Mauch and Dixon, 2010) and a criterion for distance thresholding (Casey *et al.*, 2008). In addition, we used the number of occurrences of each leitmotive during the 10-minute listening passage as a predictor for the regression model for 'old motives'.

Having only 10 observations per model, we found it necessary to reduce the number of predictor variables using stepwise backward and forward model selection using the Bayesian Information Criterion (BIC) as the model fit index, rather than using a threshold of statistical significance. The coefficients of final regression model for the 'novel' motives are given in Table 1. The model has an adjusted R^2 of 0.35 but fails to reach significance overall ($F_{(2,7)} = 3.4$, $p = 0.09$). The model includes the chroma feature distance as a predictor, indicating that motives with a large chroma distance (loosely, "sounding dissimilar") to any segment within the 10-minute listening passage are easier to identify as novel than motives with a small chroma distance (closer harmonically). In addition, the participants' valence ratings are selected as a predictor in the final model, albeit with a non-significant coefficient estimate. Here, motives rated as rather sad were more difficult to identify as novel motives. Neither the number of occurrences in the listening passage nor the perceived emotional arousal of the leitmotive were predictors in the regression model.

p-value	t statistic	error	estimate	
0.0356*	2.597	1.5628	4.0578	intercept
0.0480*	-2.392	1.5788	-3.7761	chroma distance
0.0705	-2.132	0.2135	-0.4550	valence

Table 1. Final regression model for 'novel' motives.

The final regression model for the 'old' motives is given in Table 2. The model has an adjusted R^2 of 0.23 and also fails to reach significance overall ($F_{(2,8)} = 3.7$, $p = 0.09$). The model includes the mean arousal ratings as the single predictor, indicating that motives that are perceived as more arousing are also recognised better as old motives. None of the other predictor variables

(number of occurrences, emotional valence, harmonic distance) featured in the final regression model for old items.

p-value	t statistic	error	estimate	
0.1005	1.856	1.4217	2.6390	intercept
0.0924	-1.911	0.2750	-0.5256	arousal

Table 2. Final regression model for 'old' motives.

3 Discussion

The decision to use an IRT approach was motivated by several factors which might generalise to similar research scenarios in empirical musicology. Firstly, we had to devise new measurement instruments for assessing very specific abilities that have not been well investigated before (*e.g.* Wagner expertise), and the IRT approach framework in general and Rasch modelling in particular provide a rigorous framework for constructing new tests as well as measuring the latent ability to perform on these tests. As a result the Wagner knowledge test and the Wagner affinity survey are now finished tools that can be used in any subsequent Wagner research; we have confirmed the specific objectivity of the Wagner knowledge test, and it should therefore generalise well to a new sample.

Secondly, the leitmotive recognition experiment had the dual purpose of measuring the ability of participants with different backgrounds to recognise leitmotives that they had been previously exposed to in the 10 min listening passage, as well as measuring the difficulty of individual leitmotives to be recognised or identified as novel. This dual aim – to gather data simultaneously about participants as well as about items of a tests – is not very common in psychological research which tends to focus on the psychological mechanisms of the participants, but is less uncommon in empirical music research that uses ecologically-valid stimuli. The IRT framework and the Rasch model that we used provide a very elegant way of generating data characterising participants and leitmotive items at the same time and within the same model.

The structural equation analysis using participants' ability coefficients demonstrates how important expertise and familiarity with a particular musical style are in order to perform well on a listening test with stimuli from this style – in fact, Wagner expertise proved to be much more important than musical expertise in order to perform well on the listening test. The SEM also showed that musical training did not (directly) correlate with specific Wagner knowledge, and Wagner knowledge can be regarded as an effective type of musical expertise that is not linked to instrumental practice.

The fact that the Rasch model from the listening test did not exhibit subgroup invariance suggests some caution in interpreting the results of the

subsequent regression analyses, and clearly both regression models suffer from the low item count of $N = 10$, as the coefficients of some model predictors did not reach the usual significance level. However, both regression models suggest that emotional processing of the leitmotives is linked to performance on the cognitive memory task, supporting the idea that cognitive and emotional processes during music listening are not separate but can significantly influence each other. In a forthcoming investigation, we aim to measure electrodermal activity and heart-rate data from listeners attending performances of Gergiev's production of the *Ring* and correlate those data with leitmotive occurrence.

We also found that for novel leitmotives harmonic distance in the acoustical signal was a predictor of their perceptual difficulty, indicating that harmonic distance can partially model a memory process that leads to the illusion of the familiar. However, this result should be replicated with a new set of leitmotive stimuli taken from a different passage, where the findings from the present study with regards to the influential predictor variables can serve as proper hypotheses. We also note, given Wagner's own theory of *Gefühlverständnis*, that it is unclear how much Wagner himself intended the listener to recognize leitmotive, and whether the greater difficulty we find associated with sadder motives is therefore more in line with his artistic intentions.

While the IRT approach has proved very useful for the analysis of our data, we note a few caveats. Firstly, IRT models require a substantial amount of data in order to be fit and to produce coefficients with acceptable confidence intervals. This is even more true for models with additional discrimination or guessing parameters. It is worth noting that the sample size of the memory experiment ($N = 100$) is at the lower bound of what is commonly recommended (de Ayala, 2009), even for simple Rasch models.

Secondly, not all psychological or empirical music research questions can be implemented as tests where correct/incorrect answers can be scored objectively. Much music psychological work investigates the appearance of musical stimuli and can ask for subjective perceptions rather than objective ability to perform a test (Kingdom and Prins, 2010). In these scenarios, IRT approaches appear to be less useful.

Finally, IRT models generally do not allow for a detailed analysis of the types of individual participants' biases. Here, techniques from signal detection theory (MacMillan and Creelman, 2005) that can distinguish between *e.g.* 'false alarms' and 'misses' allow for a greater insight into the nature of the cognitive processes behind the performance on a particular test and into potentially interesting interactions between both person and item characteristics.

In sum, IRT is most useful when the main research questions target individual differences between participants and data from a large sample with good variation in test performance and related background variables can be obtained. Using an IRT approach we have been able to show how individual differences in musical training and Wagner expertise lead to differential per-

formance on the leitmotive recognition task. Because recognising leitmotives in the constant auditory stream of Wagner's music affects a listener's musical perception, the individual differences we have identified may well influence the experience of Wagner's music, both in cognitive and emotional terms.

3.1 Acknowledgments

This work was supported by the *Transforming Musicology* project, AHRC AH/L006820/1.

4 References

ALBRECHT, H. (2012): Wahrnehmung und Wirkung der Leitmotivik in Richard Wagners Ring des Nibelungen – Eine empirische Studie zur Wiedererkennung ausgewählter Leitmotive aus musikpsychologischer und musiksemiotischer Perspektive. Masters Dissertation.

DE AYALA, R. J. (2009): Theory and practice of item response theory. Guilford Publications.

BIRNBAUM, A. (1968): Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord and M.R. Novick (eds.), Statistical theories of mental test scores (pp. 395-479). Reading, MA: Edison-Wesley.

BORSBOOM, D. (2006): The attack of the psychometricians. Psychometrika, 71:3, 425-440.

CASEY, M., RHODES, C. and SLANEY, M. (2008): Analysis of Minimum Distances in High-Dimensional Musical Spaces. IEEE Transactions on Audio, Speech and Language Processing, 16:5, 1015–1028

DALHAUS, C. (1979): Richard Wagner's music dramas. Cambridge: Cambridge University Press.

DELÈIGE, I. (1992): Recognition of the Wagnerian Leitmotiv. Jahrbuch der Deutschen Gesellschaft für Musikpsychologie, 9, 25-54.

GLAS, C. A. W., and VERHELST, N. D. (1995). Testing the Rasch model. In G. H. Fischer and I. W. Molenaar (eds.), Rasch models: Foundations, recent developments, and applications (pp. 69-95). New York: Springer-Verlag.

HORN, J. (1965): A rationale and test for the number of factors in factor analysis. Psychometrika, 30:2, 179-185.

KINGDOM, F., and PRINGS, N. (2010): Psychophysics: a practical introduction.

LORD, F. M. (1980): Applications of item response theory to practical testing problems. Routledge.

MACMALLIN, N. A., and CREELMAN, C. D. (2005): Detection Theory: A user's guide.

MAIR, P., and HATZINGER, R. (2007): eRm: Extended Rasch Modeling. R package version 0.15-4.

MAUCH, M. and DIXON, S. (2010): Approximate Note Transcription for the Improved Identification of Difficult Chords. In: Proc. International Society for Music Information Retrieval Conference, Utrecht, Netherlands, 135–140

MÜLLENSIENFEN, D., GINGRAS, B., STEWART, L., and MUSIL, J. (2011): The Goldsmiths Musical Sophistication Index (Gold-MSI): Technical Report and Documentation v0.9. London: Goldsmiths, University of London. URL: `http://www.gold.ac.uk/music-mind-brain/gold-msi`.
MÜLLENSIEFEN, D., and HALPERN, A. (2014): The role of features and context in recognition of novel melodies, Music Perception, 31:5, 418-435.
RASCH, G. (1960): Probabalistic models for some intelligence and attainment tests. Danmarks paedagogiske institut.
REVELLE, W. (2014): psych, Procedures for Personality and Psychological Research, R package version 1.4.8.
REVELLE, W. and ROCKLIN, T. (1979): Very Simple Structure - alternative procedure for estimating the optimal number of interpretable factors. Multivariate Behavioral Research, 14:4, 403–414.
RIZOPOULOS, D. (2006): ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses, Journal of Statistical Software, 17:5, 1-25.
SAMEJIMA, F. (1969): Estimation of latent ability using a response pattern of graded scores. Psychometrika monograph supplement.
VELICER, W. (1976): Determining the number of correlations components from the matrix of partial. Psychometrika, 41:3, 321–327.
YVES, R. (2012): lavaan: An R Package for Structural Equation Modeling. Journal of Statistical Software, 48:2, 1-36