Proceedings of the 3rd International Conference on Music & Emotion (ICME3), Jyväskylä, Finland, 11th - 15th June 2013. Geoff Luck & Olivier Brabant (Eds.)

THE EMOTIONALITY OF SONIC EVENTS: TESTING THE GENEVA EMOTIONAL MUSIC SCALE (GEMS) FOR POPULAR AND ELECTROACOUSTIC MUSIC

Athanasios Lykartsis, Andreas Pysiewicz, Henrik von Coler, Steffen Lepa

Audio Communication Group, Technical University Berlin, Germany
alykartsis@mailbox.tu-berlin.de

Abstract

In the present study the Geneva Emotional Music Scale (GEMS-25) and its German offshoot, the GEMS-28-G were tested for measurement invariance across different types of musical stimuli. Additionally, the comparability of scores across the different language versions was checked. While alternative scales are often based on general dimensional or categorical emotion theories and are thus "stimulus-neutral", the domain-specific likert-type emotion scale GEMS is designed to especially capture the emotions evoked when listening to music. Within the study, an online survey was administered (n = 245) using a stimuli set of 20 excerpts from musical pieces. By analyzing the data with structural equation modeling (SEM), we tried to verify the reliability of the scales in terms of measurement invariance towards popular/classic music as well as towards the genre of electroacoustic music, employing the latter as an extreme case of a "non-conventional musical style". We subsequently also tested for measurement invariance across languages. Concerning music styles, measurement invariance of the original GEMS-25 was achieved only at the "configural level", while the GEMS-28-G could reach at least "weak factorial invariance". This demonstrates that only for the German version the contextual meaning of the construct remains constant across different musical genres with a reasonable fit. Nevertheless, researchers should be cautious when comparing GEMS factor scores achieved with very heterogenic musical styles in future studies, regardless in which language.

Keywords: emotion measurement, structural equation modeling, measurement invariance

1. Introduction

The measurement of music-induced emotions is a growing field within music psychology, with a multitude of basic questions to be answered. A debate is still going on concerning the nature of the emotions related to music. Some claim that music does not evoke emotions as experienced in everyday situations, but that listeners respond affectively to music (Hunter & Schellenberg, 2010).

Different approaches are available for measuring emotions in general, including the measurement of bodily functions, the use of real-time user responses or surveys based on self-reports (Desmet, 2003). The latter have a long tradition in the capturing of everyday sensations. Some authors express uncertainty that theories developed for non-music-related emotion measurement are adequate for use in music related questions and its aesthetic context (Vuoskoski & Eerola, 2011, p. 160). Although the two-dimensional circumplex model and the discrete emotion model have been extensively used (ibid.), the need for a music specific instrument describing more aesthetical contexts has become obvious (Zentner, Grandjean & Scherer, 2008). This is due to the fact that emotions related to the listening of music seem to differ from those we experience in everyday situations.

To overcome the shortcomings of generic scales in music emotion research, the *Geneva Emotional Music Scale* (GEMS) has been introduced by Zentner, Grandjean, and Scherer in 2008. In a series of experiments, 45 items were selected to measure the perceived feelings. The items are organized in nine factors and three super factors. The resulting scale - as well as shorter adaptions containing 25 and nine items - was evaluated in further tests and the model showed a significant fit for classic music pieces (ibid.). Making it available in an international context, the original French GEMS was adapted to English language by Zentner and co-researchers. For a research project in German language, an adaption of the GEMS-25 has been conducted by Lepa (in preparation), namely the GEMS-28-G. Being rather new measurement tools, both scales, the English and the German, are still in the need of being tested regarding their validity and reliability to either affirm their suitability or disclose their weak spots.

Typical test designs for measuring musically induced emotions apply generic musical stimuli, mainly well-established music genres like popular and classical music. Since the GEMS is designed to capture the emotions induced by music in general and without regarding the genre, the results of an experiment with non-conventional music are crucial for defining the limits of the measurement instrument. Vuoskoski and Eerola conducted a test of the GEMS using "*ecologically valid and emotionally diverse stimulus material*" (Vuoskoski & Eerola, 2011, p. 160) and could demonstrate the suitability of the scale for music in a wider sense. Since music as a cultural technique is always in constant change (especially during the 20th century, the boundaries of what is considered as music were redefined), it is of interest whether tools for measuring musical emotions can account for these changes.

As a consequence, we focused on testing for scale invariance with different musical stimulus material (for the concept of measurement invariance see Section 3), in particular using the genre of electroacoustic music[1]. Electroacoustic music clearly marks the boundary between what is regarded as music in the common sense and sound art, which would not be categorized as music if listened to by laymen.

It is of fundamental interest, whether the emotions evoked when listening to avantgarde music can be measured in the same way as those evoked from popular music genres. Thus the experiment presented in this paper can be regarded as one of many experiments defining the genre limits, in which a music specific emotion measurement is reasonable.

Our experiment aimed at two different research questions, namely the testing for invariance and of the respective model fit for the GEM scale for (1) different stimulus material (i.e. popular/classical music and electroacoustic music) and (2) over independent test groups using the English and the German versions of the measurement scale. Consequently, within our study measurement invariance was tested across measurement occasions (stimuli) and, in a second step, across groups (languages) at a common set of hierarchically structured levels of invariance constraints (Widaman & Reise, 1997): (a) configural invariance, (b) weak factorial invariance (equal factor loadings), (c) strong factorial invariance (equal factor loadings and intercepts) and (d) strict factorial invariance (adds equal error residuals). The degree of invariance defines the premises for comparisons between resulting parameters of the scale (eg. factor means) when applied to different measurement occasions and the conclusions that can be drawn from them.

This paper is further organized as follows: In Section 2 the GEM scale as well as the test environment and procedure are explained in detail. The evaluation procedure and related methods are presented in Section 3. Results are presented in Section 4 and discussed in Section 5. The paper ends with a final conclusion in Section 6.

[1] *The term electroacoustic music refers to a modern form of Western art music and compositional practice that makes use of electric sound (re-)production. Therefore it includes tape music, electronic and computer music. For further information on electroacoustic music refer to Böhme-Mehner, Mehner & Wolf (2008).*

2. Experimental Setup

2.1. Measurement

In our present experimental setup we made use of the German translation of an extended[2] GEMS-25, the GEMS-28-G as introduced and evaluated by Lepa (in preparation). This scale consists of 27 labels (adjectives) representing the nine categories of GEMS-9 with three items per dimension. These dimensions condense into three top-level factors (cf. Table 2).

In order to address our research question adequately, we made the survey accessible to non-German speakers by using the English version consisting of equivalent terms of the extended GEMS-25 on which the German translation was based on. In this context, it has to be considered that shifts in meaning of terms may occur, as nuances of the original scale could have been "lost in the translation" as pointed out by Vuoskoski and Eerola (2011), who developed a Finnish translation of the English attributes.

The item batteries containing the German and the respective English adjectives can be found in Table 2. For rating the intensity of the given items, the original 5-point (Likert) scale ranging from 1 (not at all) to 5 (very much) was used (cf. Table 1). The respective scale was designed to be balanced, non-forced choice and even interval scaled.

2.2. Procedure

Our approach (being based on a within-subjects design) was to present two excerpts (one of each stimuli group) to the participants to be evaluated in terms of induced, felt emotions on the provided rating scale. In order to make the test available for a large number of participants and to avoid the effort of evaluating paper questionnaires, we decided to use the online survey application *LimeSurvey* that was adapted to our specific requirements. An audio player was implemented to play back the sound excerpts within the survey. Moreover, both the stimuli and the respective item list for each sound excerpt had been randomized before presented to the individual listener to avoid any kind of systematic sequential effects and response biases. Furthermore, the two musical excerpts to be presented to the individual subject were also chosen randomly resulting in 100 (= 10^2) different stimuli combinations.

The testing procedure within the online survey was structured in five parts being presented on consecutive screens:

(1) *Introduction:* Participants were instructed regarding the testing procedure and their task to rate the felt emotions in context of the presented music.

(2) *Sound system setup:* For optimized listening conditions, we provided an audio setup consisting of an audio player with a *neutral* musical excerpt[3] and instructions to adjust volume, eliminate noise sources and get into a comfortable listening position.

(3 - 4) *1st and 2nd music excerpt:* In these parts the auditory stimuli were rated according to their affectional impact, in other words, the intensity of the felt emotions on basis of the emotion labels provided by the GEM scale.

(5) *Personal questions on participants*: In order to allow the statistical analysis of the sample data, the participants were asked to provide sociodemographic information and to describe their relationship to music and individual music listening habits.

2.3. Stimuli

As auditory stimulus material we used two categories of music excerpts: The first group, labelled as *anchor music*, consisted of ten excerpts of purely instrumental popular and classical music. The pieces were chosen to represent a variety of ordinary music genres most listeners should feel familiar with. The second group of stimuli was a compilation of ten different pieces of electroacoustic music. Choosing the excerpts, the goal was to represent a spectrum as large as possible of respective

[2] *Two adjectives from the original extended GEMS-45 version (cf. Section 1), "blue" and "nervous" were added to the GEMS-25 scale for achieving comparability.*

[3] *A song excerpt representing a bright mood and being normalized to a peak level of -2 dB with respect to the test stimuli to be rated in order to add slightly more presence to the latter.*

musical aesthetics, moods and production techniques.

All excerpts from both groups had a duration of two minutes with a soft fade-out at the end of each excerpt.

The sound excerpts were encoded with the lame mp3 codec (320kBit/s, 44.1 kHz, Stereo) in order to reduce file size and to ensure streaming via the Internet (with a negligible loss of audio quality).

2.4. Participants (n = 245)

The survey was spread via different mailing lists and online interest groups, related to musicology, sound art as well as non music specific groups. In this way, we were able to reach a target audience beyond a sample consisting exclusively of students being involved with the specific problem. Altogether, 245 participants took part in the survey. More than half of them (55.5%) were German speakers, 45.5% participated in the English version of the survey. Around 57% of all participants were male; the average age of the samples (ranging from 15 to 71 years) was 28.5 years (SD = 7.5). Regarding their educational level, almost two-thirds (63.3%) of the participants stated to have a higher education degree, a further 31.4% graduated from college. Concerning musicological knowledge, the largest group of 42.5% indicated having a good understanding of music, while laymen and music experts comprised each 28.6%. Their average music listening time was 110.7 minutes (SD = 99) per day. In matters pertaining to the experiment, 88.5% of all participants rated the testing conditions as good or even very good, only 2.2% reported having bad conditions of participation.

3. Method

Following the method of data evaluation in Zentner et al. (2008), the acquired data underwent a confirmatory factor analysis (CFA) The experimental questions to be answered can be summarized as follows: Is the measurement tool invariant across different musical stimuli and, if so, to what extent? Furthermore, can measurement invariance be observed across groups of different languages?

The test results allow conclusions concerning the reliability of the model over a wide range of possible musical contexts and stimuli, in order to further evaluate the method proposed by Zentner et al. (2008).

3.1. The Factor Model

The basis of the analysis is the factor model shown in Table 2b. It specifies the loading structure of the 27 adjective labels, the items, in nine first-order latent variables (factors), which themselves load onto three second-order factors. This model tries to explain the variability and correlation observed in the item covariance matrix based on the assumption that they are caused by fewer, unobserved variables (the factors) whose scores are dependent on a linear combination of the item scores, their likewise difficulty (i.e. the item intercept value) and an item-specific error term (Backhaus, Erichson, Plinke & Weiber, 2006). Each of the items is associated with a specific factor loading, which shows the extent to which the property referred to by the observed variable (in this case the adjective characterizing the felt emotion) contributes to the concept described by the higher order factor (here a mood, disposition or emotional state).

3.2. Measurement Invariance

Being confronted with the issue whether the measurement tool provides reliable results for different kinds of stimuli and independent groups of test subjects, testing for measurement invariance is the procedure of choice (e.g. Geiser, 2010; Widaman & Reise, 1997; Meredith, 1993). It involves testing for identical (equivalent) constructs across measurement occasions or groups to assure a comparability of measures. On this premise, the same constructs are being assessed and a meaningful comparison of statistics (such as means and variances) of each measurement can be performed. As mentioned earlier, measurement invariance can be tested at different levels representing increasing measurement model constraints of a nested model set. In our work, we followed the hierarchical set of model tests described by Widaman and Reise (1997), consisting of four levels of invariance constraints:

(a) The first, basic level of measurement invariance is called *Configural Invariance*. As its name implies, it refers to the same configuration of factor loadings. The factor loadings can differ for each measurement occasion or group. When this level of nonmetric invariance is met, the latent variables, which are present within each construct, are similar, but not identical (Widaman & Reise, 1997, p. 292).

(b) The second level of invariance (and first form of a metric factorial invariance) is called *Weak Factorial Invariance*. In addition to the requirement of configural invariance, each item's loading to the respective factor (indicating the strengths of the linear relation between a factor and the associated items) is restrained to be equal for each measurement occasion. Accordingly, with constant factor loadings the scale unit is identical in each construct, though the scale origins are not necessarily the same.

Table 1. 5-point rating scale (both in German and in *English*)

1	2	3	4	5
überhaupt nicht	eher nicht	mittelmäßig	ziemlich	sehr stark
not at all	*not really*	*more or less*	*quite*	*very much*

Table 2a & b. Emotional labels with first and second-order factors (both in German and in English)

German		
Items	**1st order factors**	**2nd order factors**
Bewegt	Bewunderung	Erhabenheit
Verzaubert		
Verträumt		
Fasziniert	Transzendenz	
Überwältigt		
Gefesselt		
Gelassen	Beruhigung	
Entspannt		
Relaxed		
Sensibel	Sensibilisierung	
Ergriffen		
Entrückt		
Nostalgisch	Nostalgie	
Sentimental		
Romantisch		
Kraftvoll	Stärke	Vitalität
Energetisiert		
Triumphierend		
Munter	Anregung	
Heiter		
Fröhlich		
Traurig	Traurigkeit	Unbehagen
Melancholisch		
Schwermütig		
Fahrig	Anspannung	
Gereizt		
Nervös		

English		
2nd order factors	**1st order factors**	**Items**
Sublimity	Wonder	Moved
		Filled with wonder
		Allured
	Transcendence	Fascinated
		Overwhelmed
		Feeling of transcendence
	Peacefulness	Serene
		Calm
		Soothed
	Tenderness	Tender
		Affectionate
		Mellow
	Nostalgia	Nostalgic
		Sentimental
		Dreamy
Vitality	Power	Strong
		Energetic
		Triumphant
	Joyful Activation	Animated
		Bouncy
		Joyful
Unease	Sadness	Sad
		Tearful
		Blue
	Tension	Tense
		Agitated
		Nervous

When this level of invariance holds, the relationship between factors and other external variables is comparable across occasions and groups. Still, no comparison of factor means is valid, due to a possibly different scale origin.

(c) The third level of invariance is *Strong Factorial Invariance.* This form of invariance postulated that not only the factor configuration and item loadings on the underlying factor are invariant over time or across groups but also the intercepts of the measured variables are equal. According to Widaman and Reise (1997, p. 294), this level of invariance is required for a meaningful comparison of mean differences of the latent variables over time or across groups. Thus, differences in factor means can be identified as true differences between measurement occasions, not being artefacts of measurement.

(d) The fourth and last level of invariance is called *Strict Factorial Invariance*. In testing this form of invariance, additionally to the three preceding constraints, the measurement residual (or measurement error) associated with each measured variable is restrained to be equal over time/across groups. When this level of invariance is achieved, all group differences are exclusively due to group differences on the common factors. Strict factorial invariance, however, is seldom found to hold for a variety of reasons (cf. Widaman & Reise, 1997, p. 296).

3.3. Evaluation

The confirmatory factor analysis was performed using the free statistics software *R* and the package *lavaan* for multivariate analysis. A matrix carrying the data for both measurement occasions (stimuli) was produced. For each language, the structural equation model which expressed the structure of the first and second-order factors with respect to the items (cf. Table 2) was defined in *lavaan*. The first step was a test for autocorrelation between the model parameters taken at the two occasions (corresponding to the different stimuli categories) of the experiment. This was achieved by gradually imposing less constraints to the Basic Model (α) - allowing uncorrelated error variances (β), uncorrelated factors between different occasions (γ), uncorrelated factors within an occasion (δ) - and observing the fit of the model and its change in respect to the prior model. The model by which the fit does not deteriorate significantly is selected as the most parsimonious and equally potent to continue to the next stage of measurement invariance testing. Such a model contains less parameters and has the advantage of being less complex and computationally costly. However, it should be mentioned that a model with more restrictions is also prone to be more difficult to fit.

We operate on a nested model where additional structural constraints were imposed in each step, as shown in the previous paragraph. To determine the level of measurement invariance present, the fit is tested again as in the case of autocorrelation by means of a likelihood ratio test (LR-test) between the fit of the current model in respect to the previous one. If this value is below a certain threshold, the two measurement models differ significantly from each other, which can also be observed by a noticeable deterioration of the fit in the case of imposing more constraints. In that case, the previous model is retained and the corresponding constraints specify the level of invariance. Finally, we test in the same way as mentioned above if the scale measures reliably between groups of speakers of different languages, in order to assert the measurement invariance of the scale for different language groups. All the tests were performed for the first-order factor model, as well as the expanded version with the second-order factors.

At this point, a short discussion about the model fit indices used in this study is necessary. The most commonly used indices for model fit comparison are the chi-square (χ^2) function value, the root-mean-square error of approximation (RMSEA), the standardized root-mean-square residual (SRMR), the Bayesian information criterion (BIC), the Akaike information criterion (AIC) and the comparative fit index (CFI). For a detailed description of fit indices and their properties see Hu and Bentler (1999), Backhaus et al. (2006), Bentner (1990) and Hooper, Coughlan and Müller (2008). As Zentner et al. (2008, p. 505) mention in the discussion part of their study, these indices can be separated into two categories based on the sensitivity towards weakly specified factor covariances (SRMR) and towards imperfect factor loadings (RMSEA, CFI and others, cf. Fan & Sivo, 2005). In this paper, we decided to use a combination of three indices as well as the chi-squared function value and two information measures (BIC, AIC) (cf. Tables 3 - 8) in order to draw conclusions about the goodness of fit.

Caution is called for absolute fit indices such as χ^2 and SRMR, as sample and model size have an influence on them that can be misleading (Hu & Bentler, 1999). However, a combination of good results for at least one of the RMSEA and CFI and the SRMR is a good indicator for a reasonable model fit.

Table 3. Results of autocorrelations and fit comparison for first-order factor model (German data)

Model (n = 136)	Par.	χ^2	df	p(χ^2)	p(LR)	RMSEA	SRMR	CFI	AIC	BIC
α) Basic Model	*288*	*2053*	*1197*	*.000*	-	*.073*	*.083*	*.791*	*20071*	*20910*
β) Non-Correlated Errors	*261*	*2088*	*1224*	*.000*	*.1479*	*.072*	*.083*	*.789*	*20052*	*20812*
γ) No Inter-Correlations	180	2295	1305	.000	< .001	.075	.137	.758	20098	20626
δ) No Intra-Correlations	108	3473	1377	.000	< .001	.106	.211	.488	21131	21446

Table 4. Results of model fit test for nested measurement models, first-order (German data)

Model (n = 136)	Par.	χ^2	df	p(χ^2)	p(LR)	RMSEA	SRMR	CFI	AIC	BIC
a) Basic Model	*342*	*2053*	*1197*	*.000*	-	*.073*	*.082*	*.791*	*20179*	*21175*
b) Eq. Item Loadings	*324*	*2075*	*1215*	*.000*	*.2432*	*.072*	*.083*	*.790*	*20165*	*21109*
c) Eq. Item Intercepts	297	2329	1242	.000	< .001	.080	.189	.734	20366	21231
d) Eq. Error Variances	243	2494	1296	.000	< .001	.082	.189	.707	20422	21130

Table 5. Results of autocorrelations and fit comparison for first-order factor model (English data)

Model (n = 109)	Par.	χ^2	df	p(χ^2)	p(LR)	RMSEA	SRMR	CFI	AIC	BIC
α) Basic Model	*288*	*2047*	*1197*	*.000*	-	*.081*	*.095*	*.740*	*15674*	*16449*
β) Non-Correlated Errors	261	2131	1224	.000	< .001	.082	.095	.723	15703	16406
γ) No Inter-Correlations	180	2333	1305	.000	< .001	.085	.160	.686	15744	16228
δ) No Intra-Correlations	108	3482	1377	.000	< .001	.118	.241	.356	16749	17040

Table 6. Results of model fit test for nested measurement models, first-order (English data)

Model (n = 109)	Par.	χ^2	df	p(χ^2)	p(LR)	RMSEA	SRMR	CFI	AIC	BIC
a) Basic Model	*342*	*2047*	*1197*	*.000*	-	*.081*	*.093*	*.740*	*15782*	*16703*
b) Eq. Item Loadings	324	2130	1215	.000	< .001	.083	.105	.720	15829	16701
c) Eq. Item Intercepts	297	2362	1242	.000	< .001	.091	.211	.658	16007	16806
d) Eq. Error Variances	243	2597	1296	.000	< .001	.096	.206	.602	16134	16788

Table 7. Results of model fit test for nested measurement models: language group comparison, first-order (popular music)

Model (n = 245)	χ^2	df	p(χ^2)	p(LR)	RMSEA	CFI	BIC
a) Basic Model	*1105*	*576*	*.000*	-	*.087*	*.856*	*18809*
b) Eq. Item Loadings	1153	594	.000	< .001	.088	.848	18758
c) Eq. Item Intercepts	1321	612	.000	< .001	.097	.808	19124
d) Eq. Error Variances	1474	621	.000	< .001	.106	.782	19227

Table 8. Results of model fit test for nested measurement models: language group comparison, first-order (electroacoustic music)

Model (n = 245)	χ^2	df	p(χ^2)	p(LR)	RMSEA	CFI	BIC
a) Basic Model	*1068*	*576*	*.000*	-	*.084*	*.816*	*18292*
b) Eq. Item Loadings	1141	594	.000	< .001	.087	.796	18266
c) Eq. Item Intercepts	1309	612	.000	< .001	.096	.740	18632
d) Eq. Error Variances	1380	621	.000	< .001	.100	.738	18654

4. Results

The test results are presented in Tables 3 - 8 (*italics* denote the models which achieved acceptable fit indices). Altogether, the achieved model fits cannot be described as satisfying, a result that can be attributed to the combination of relatively small sample size (regarding this kind of analysis) and high model complexity. However, this does not constitute a problem for the study at hand, as the absolute values of the fit indexes are not of importance for the investigation of measurement invariance, only their relative differences between the nested models.

The autocorrelation test showed that for the German version the change in model fit is negligible when the error correlations are set to zero. Therefore these parameters could be let aside, creating a less computationally costly and simpler model. However, we discovered that the improvement was only marginal (a positive change in fit indices of 1% in the test for measurement invariance and the ability of conducting alikelihood ratio test actually impaired, for both the first and the second-order factor model. We therefore used the basic model configuration (all parameters correlated). For the English version of the scale the autocorrelation test was not conclusive, therefore we retained the original model with the maximum amount of correlations in that case as well.

The test for measurement invariance provided similar results: For the German dataset, weak factorial invariance could be demonstrated, for both the first and the second-order factor model. For the English version only configural invariance was attested. One should note, however, that the overall fit in all cases was too low (CFI < 0.8) to meet the formal requirements for a good fit as proposed in Hu & Bentler (1999). Nevertheless, the German data set provided overall better fit indices than the English one (e.g. $\text{CFI}_{\text{German,1st-order}}$ 0.791 vs. $\text{CFI}_{\text{English, 1st-order}}$ 0.740 for the basic model, respectively for all other fit indices cf. Table 4 vs. 8).

Finally, the tests for measurement invariance across the language groups returned significant values for the LR-test in all cases (see the column under p(LR) in Tables 3-8), thus allowing to ascertain only configural invariance. In total, the results might be summed up as follows: for the first and second-order factor model, configural invariance is present across both stimuli and languages, whereas weak factorial invariance is only present in the case of the German scale across measurement occasions (stimuli). The results concerning the second-order factor model are not presented here, as the attained invariance levels are identical to those related to the first-order factor model and the index values do not differ substantially (cf. Section 5).

5. Discussion

Considering the results in Section 3, several major issues arise. Firstly, the overall fit index values and results of the likelihood ratio test were substantially better in the case of the German data (cf. e.g. Table 4 vs. 6). These differences can be traced back to the increased sample size (n = 136 for German, n = 109 for English) and ratio of amount of participants to items (5:1 for German, 4:1 for English). The amount of samples in the present study was sufficient for conducting a confirmatory factor analysis in all cases (Gorsuch, 1983; Fan, Thompson & Wang, 1999). However, the general rule holds that the greater the sample size, the more accurate the statistical evaluation, which in our case explains the poorer performance in the English language case.

With respect to the different language versions, our results also suggest that the German version of the items list is semantically closer to the original, French adjective model of Zentner et al. (2008). This might have resulted in a better comprehension of the adjective meaning, leading to the production of comparable results to the aforementioned study for the German version. Furthermore, with regard to the English version, only 9.2% of the participants in this case could be assumed to be native speakers (having accessed the online survey from Australia, Ireland, the UK and USA). It is debatable if the rest of the participants, presumably non-native speakers, could grasp the fine nuances in meaning between the adjective items, resulting in a systematic error in that case. This becomes apparent when observing the results of the measurement invariance test for the two language groups, which attested only configural invariance (cf. Table 7 and 8).

Another important point of the present study diverging from the work by Zentner et al. (ibid.) is the difference in test stimuli and test conditions. The pop music repertoire included musical excerpts from many different genres, whereas within the reference study basically classical music was featured. This could explain the overall lower fit indices observed in our case as the original items are presumably more appropriate for emotions evoked by classical music. Apart from that, the researchers used a listening context of live performance in their research, which lead to possible differences in nature and magnitude of felt emotions with respect to the listening situation in our study, as indicated as an important *contextual feature* by Scherer and Zentner (2001, p. 364).

Moreover, it should be mentioned that fit indices for the second-order factor model were inferior to those of the first-order model as the increased complexity makes it more difficult to fit. Yet, the results concerning the degree of measurement invariance are consistent between the respective first and second-order factor model, which adds to the general solidity of the factor model proposed by Zentner et al. (2008).

Concerning the main methodological question, the attainment of weak measurement invariance for the German data can be described as satisfactory if a study aims solely at analysing covariance structure models, since in this case it holds to compare variances or covariances between the latent variables and external variables such as, e.g. age and gender (as factor means are of no interest in this case). It is, however, not feasible to compare the factor means (and therefore neither the ratings of the stimuli can be compared nor conclusions about changes in factor scores over time can be drawn).

The absence of strong factorial invariance in all cases (over time and across groups) is caused by consistently different item intercepts in the case of electroacoustic music as compared to popular music. This result might be construed as an indication that participants tend to evaluate the former in a different way than popular music genres, suggesting that some item indicators might be superfluous or misleading. The presence of configural invariance in the case of English data shows that the model is not strictly invariant across stimuli, but that the measured emotions tend to have the same character ensuing from an equal factor configuration. The contextual meaning of the factor construct remains the same across stimuli. The factor loadings however are free to vary, implying that in this case the relationship between felt emotions and the factors to which they belong is not stable over stimuli.

6. Conclusions

Our study confirmed that the GEM scale may be used to measure invariantly across stimuli and languages at the configural level with a reasonable fit. But, as pointed out in section 4, configural invariance does not attest a high degree of reliability of the measurement scale. Thus, it only confirms the suitability of the model proposed by Zentner et al. (2008) for music induced emotion measurement on a basic structure level. Based on our data, comparing variances and covariances across measurement occasions should only be performed in the case of the German version. The reasons for this result can be traced back to the configuration of the scale, the different listening context or the lack of familiarity of the participants with these stimuli. More concrete results could be achieved by using excerpts with more specific character or musical form, or being representative of a specific mood. In this way it could be determined if the emotions evoked by electroacoustic music have an affinity to those evoked from very specific musical genres. Another approach would be testing for measurement invariance for subgenres of ordinary music in order to determine the scope in which the scale does measure invariably. Furthermore, it might be possible that modification of the model (e.g. by applying an explorative factor analysis in order to detect which items and factors can be excluded) could yield a factor structure which would serve as a common measurement instrument for a wider range of music styles. In this case, an extended research as to which items or factors could be modified should be undertaken, as well as new tests should be conducted to confirm the results. Such an approach is out of the scope of our study, but could serve as a starting point for future experiments.

References

Backhaus, K., Erichson, B., Plinke, W., & Weiber, R. (2006). *Multivariate analysemethoden: eine anwendungsorientierte einführung* (Vol. 11). Berlin: Springer.

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin*, *107*(2), 238 - 246.

Boehme-Mehner, T., Mehner, K., & Wolf, M. (2008). *Electroacoustic music - Technologies, aesthetics, and theories - A musicological challenge*. Essen: Blaue Eule.

Desmet, P. (2005). Measuring emotion: development and application of an instrument to measure emotional responses to products. In: *Funology* (pp. 111 - 123). Springer Netherlands.

Fan, X., Thompson, B., & Wang, L. (1999). The effects of sample size, estimation methods and model specifications on SEM fit indices*. Structural Equation Modelling: A multidisciplinary Journal*, *6*(1), 56-83.

Fan, X., & Sivo, S. A. (2005). Sensitivity of Fit Indices to Misspecified Structural or Measurements Model Components: Rationale of Two-Index Strategy Revisited. *Structural Equation Modeling*, *12*(3), 343-367.

Geiser, C. (2011). *Datenanalyse mit Mplus: eine anwendungsorientierte Einführung.* Wiesbaden: VS-Verlag für Sozialwissenschaften.

Gorsuch, R. L. (1983). *Factor Analysis.* (2nd. Ed). Hillsdale, NJ: Erlbaum.

Guilford, J. P., (1954). *Psychometric methods*. (2nd edition). New York: McGraw Hill.

Hooper, D., Coughlan, J., & Mullen, M. R., (2008). Structural Equation Modelling: Guidelines for Determining Model Fit. *The Electronic Journal of Business Research Methods*, *6*(1), 53-60.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55.

Hunter, P.G., & Schellenberg, E.G. (2010). Music and Emotion. In *Music Perception* (pp. 129–164). New York: Springer.

Sloboda, J.A., & Juslin, P. N. (Eds.). (2001). *Music and emotion: theory and research.* Oxford University Press.

Lepa, S. (in preparation). GEMS-28-G. Unpublished reference paper, TU Berlin, Audio Communication Group.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance*. Psychometrika*, *58*(4), 525-543.

Scherer, K. R., & Zentner, M. R. (2001). Emotional Effects of Music: Production Rules. *Music and emotion: theory and research*, 361-392.

Suhr, D. D. (2005). Principal component analysis vs. exploratory factor analysis. *SUGI 30 Proceedings*, 203-230.

Vuoskoski, J. K., & Eerola T. (2011). Measuring music-induced emotion: A comparison of emotion models, personality biases, and intensity of experiences. *Musicae Scientiae*, *15*(2), 159-173.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain*. The science of prevention: Methodological advances from alcohol and substance abuse research*. 281-324.

Zentner, M., Grandjean, D., & Scherer, K. R. (2008). Emotions Evoked by the Sound of Music: Characterization, Classification, and Measurement Emotion, 8(4), 494-521.