sempre: Society for Education, Music and Psychology Research
MUSIC & SCIENCE

Research Article

Music & Science
Volume 9: 1–17
© The Author(s) 2026
DOI: 10.1177/20592043251413550
journals.sagepub.com/home/mns
Sage

Measuring Emotional Experiences with Music: Content Validity Assessment for Episode Model Constructs

Connor Kirts[1], Suvi Saarikallio[2], Cameron J. Anderson[3], Scott Bannister[4], Julian Céspedes-Guevara[5], Gladys J. Heng[6], Noah Henry[7], Kelly Jakubowski[1], Friederike Koehler[2], Amanda E. Krause[8], Thomas M. Lennie[9], Isabel Cecilia Martínez[10], Katherine O'Neill[11], Lindsay Warrenburg[12] and Tuomas Eerola[1]

Abstract

This study focuses on the first stage of instrument development, content validity, and provides guidance on what steps should be taken when designing and evaluating content through the development of an instrument to reflect a recent theory concerning emotional episodes. To establish this instrument, we (1) operationalized the theory, identifying 25 sub-constructs for topics such as listening attention, meaning generation, preferences, familiarity, reward, and functions attributed to the use of music to regulate affectual states; (2) proposed a set of items (N = 495) to represent the relevant constructs operationalized here from the Episode Model; (3) analyzed and reduced the item pool with natural language processing (NLP); (4) assessed whether items were indeed reflective of their assumed construct and would garner appropriate responses using feedback supplied by subject matter experts; (5) refined the item pool based on expert feedback; (6) reassessed the revised items which resulted in a reduced item sample (N = 168). Through this collaborative content validity process, experts supported the theoretical positioning of the Episode Model through their agreement with the operationalized constructs. Expert insight shaped the implementation of theory from loosely associated items into a more tightly interrelated set of items, comprising fewer and more distinct constructs. We conclude by discussing the purpose of these content validity processes and outline the next stages of instrument development to construct a robust instrument which contextualizes emotional episodes experienced with music.

[1] Department of Music, Durham University, Durham, UK
[2] Centre of Excellence in Music, Mind, Body and Brain, Department of Music, Art and Culture Studies, University of Jyväskylä, Jyväskylä, Finland
[3] Department of Psychology, Neuroscience & Behaviour, McMaster University, Hamilton, Canada
[4] School of Music, University of Leeds, Leeds, UK
[5] Departamento de Estudios Psicológicos, Universidad Icesi, Cali, Colombia
[6] Centre for Music and Health, Yong Siew Toh Conservatory of Music, National University of Singapore, Singapore
[7] Institute for Logic, Language and Computation, University of Amsterdam, Amsterdam, the Netherlands
[8] Department of Psychology, James Cook University, Townsville, QLD, Australia
[9] Philosophy and Psychology Department, American University in Bulgaria, Blagoevgrad, Bulgaria
[10] Laboratory for the Study of Musical Experience, Facultad de Artes, Universidad Nacional de La Plata, La Plata, Argentina
[11] School of Arts and Creative Technologies, University of York, York, UK
[12] Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA

Corresponding Author:
Connor Kirts, Department of Music, Durham University, Durham, DH1 3RL, UK.
Email: connor.g.kirts@durham.ac.uk

Data Availability Statement included at the end of the article

Keywords
Affect regulation, content validity, context, emotion, functions of music, measurement

Submission date: 16 July 2025; Acceptance date: 6 December 2025

Despite growing literature on how music evokes emotions, there is considerable disagreement about how to explain this phenomenon (Warrenburg, 2020). Such divergence impacts what methodological choices are made to capture this phenomenon, making it difficult to compare findings between studies. A multitude of variables concerned with the music, the listener, and the context are agreed to influence the induction of emotion through various components (e.g., physical, affectual, appraisal, and functional; see Juslin & Laukka, 2004; Scherer & Zentner, 2001). Influential music–emotion models have predominantly focused on explaining interactions between musical features and listeners' conscious or unconscious induction of emotion (Juslin & Västfjäll, 2008; Zentner et al., 2008). Over time, scholars have highlighted how theories also need to address functional uses of music (Baltazar & Saarikallio, 2016; Saarikallio, 2012) and the inclusion of contextual information (Céspedes-Guevara & Eerola, 2018; Cespedes-Guevara, 2023). Some have sought to explain interactions between the listener's association with the music in context with the addition of goal-directed appraisals (Lennie & Eerola, 2022; Scherer & Coutinho, 2013). However, theorizing should seek to go beyond explaining interactions between the music, the listener, and the context to account for the procedural nature of emotions as dynamic self-regulatory episodes.

The Episode Model (Eerola et al., 2025) has recently integrated functional uses of music and contextual variables in an effort to situate and contextualize how these experiences emerge. In this model, episodes are defined as situated affectual states that emerge from an interrelated collection of subevents toward an object (Russell & Barrett, 1999). Emotional episodes are experiences of emotion, such as the joy upon seeing an artist walk on stage, terror during a first dance, awe at being immersed in a soundfield, or calmness while listening to an album. The Episode Model seeks to explain how – even if the same person listened to the same piece of music in all of these examples – emotions would change to reflect different functional uses, often directed by the listener's attention and modified by the meaning attributed to the music (Eerola et al., 2025). The Episode Model supports a more dynamic understanding of music-induced emotions, where goal-driven and stimulus-driven processes may interact (van Ede et al., 2020). Considering that vast portions of this theory are currently untestable without a formulated self-report instrument, we plan to operationalize necessary latent constructs (hereafter constructs; for a definition, see Hoyle et al., 2025[1]) from the Episode Model according to established psychometric practices.

Psychometric instruments, such as self-report questionnaires, should have four essential qualities: reliability, validity, discriminatory power, and extensive norms (Kline, 2000). Researchers typically focus on the reliability of a test (referring to the consistency of scores from one assessment to another, see Cook & Beckman, 2006, p. 13) to gauge its quality. However, validity (referring to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests; see American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 2014, p. 11) carries more substantial implications. Researchers should be able to judge whether a test is suitable for their purposes from the evidence that test developers used to substantiate an instrument.

In contemporary practice, scholars commonly employ a "one-size-fits-all" approach to test development that does not provide the rigorous tests necessary to falsify hypotheses concerned with the validity of a test but nonetheless *appear* to produce favorable results because they are reliable (Borsboom, 2006; Maul, 2017). For instance, it is common to conflate estimates of internal consistency as a sole indicator of validity (Cronbach $\boldsymbol{\alpha}$, Cronbach, 1951; e.g., Chamorro-Premuzic & Furnham, 2007; Kreutz et al., 2008; Mas-Herrero et al., 2013) instead of substantiating a test's validity through multiple appropriate investigations – concerned with its content, participant response processes, internal structure, relationships to other variables, and consequences of testing – for a desired purpose. Another common practice forgoes formulating theory to underpin a test's development and instead relies on post hoc or ad hoc explanations of factor structure to inform conceptual positions (e.g., Sandstrom & Russo, 2013; Zentner et al., 2008). Some instruments are made explicitly following such procedures for practical reasons. Zentner and colleagues (2008), for example, employed an inductive procedure to generate their items, which consequently informed the factors of their model, classifying different musically evoked emotions. However, as Maul (2017) demonstrated, it is more difficult to justify the inclusion of content and falsify dimensionality without being able to test hypotheses – in congruence with theory – to support whether responses to these items have any meaning to the supposed construct. Another consequence of insufficient theory includes making assumptions about the interactions between variables. As Borsboom (2006) noted in general psychology research, it is often concluded that observed scores indicate differences between participants prima facie. Leading test developers and researchers alike to determine correlations between a test score and a criterion as favorable by "eyeballing" the correlation matrixes as comparable instead of conducting tests of measurement invariance for example (see Millsap, 2007).

Popularity of these practices highlights how test developers are often influenced by metrics of internal consistency

and parsimony of fit rather than seeking to establish a measurement's validity with various pieces of validity evidence, informed alongside a theory of the construct (Borsboom, 2023; Revelle, 2024). Rare exceptions in music psychology have examined multiple forms of validity evidence during their test development process (see, for example, Groarke & Hogan, 2018; Henry et al., 2024; Krause et al., 2020; Law & Zentner, 2012; Müllensiefen et al., 2014), however even these have not transparently reported how items were generated or evaluated to reflect their constructs. Such practice implies that our understanding of what a test reports to measure – or even what it consists of – is vague at best and completely missing at worst. The opportunity persists to formulate instruments that have undergone transparent, rigorous, and systematic development – starting from theory, to inform and investigate the capacity of a set of items to produce valid responses – rather than assuming synonymy between the observable and latent behavior.

Research Objectives

We sought to generate items that represent the theoretical constructs from the Episode Model (Eerola et al., 2025), as the basis for a self-report questionnaire that can be used in future empirical research. Here, we establish what content is suitable to reflect our conceptualization of the theoretical constructs and investigate whether we have sufficiently sampled items to reflect these through specified content domains. We organize the text into three sections to provide an overview of methods used to conduct different steps of the content validity process, as we detail the development of our instrument to capture constructs for the Episode Model, illustrated in Figure 1 (B); (1) we define constructs from theory, operationalizing these into plausible content domains for sub-constructs, (2) we articulate how items were generated to represent sub-constructs, either adapting items from existing measures or creating them, and (3) we assess whether items are relevant and representative with evaluation and feedback from subject experts.

Our primary aim is to implement operational definitions of the relevant constructs outlined by the Episode Model and assess how well items represent the content of the constructs by using subject experts. Our secondary aim is to gain insight into existing items for constructs commonly used in music psychology research. However, we will not provide a holistic review of current best practice for reporting measurement choices (see Flake & Fried, 2020); instead we wish to highlight this often-overlooked stage of test development to discuss and improve the field's creation of content to reflect constructs used in empirical research.

Defining Constructs from Theory

Music psychologists and related scholars have not shied away from developing psychometric instruments. A recent scoping review found 56 psychometric instruments, used to quantify music-related constructs, alone (Koehler et al., 2025). Considering these tools are used as the basis for empirical research, it should be reasonable to assume that rigorous standards have been used when developing, analyzing, and scrutinizing the validity of these tools. Such obligations are stated in the *Standards for Educational and Psychological Testing,* where validity is considered "the most fundamental consideration in developing and evaluating tests" (AERA et al., 2014, p. 11). In consequence, providing evidence for the validity of a test and its usefulness for empirical hypothesis testing depends on the extent to which it is based on a theory (Borsboom, 2023, 2006; Epskamp et al., 2017). Multiple scholars reiterate that the first stage of psychometric development involves articulating a framework to support the claim that specified content is capable of reflecting a theoretical construct (Borsboom, 2023; Sireci, 1998; see also Cronbach & Meehl, 1955; Loevinger, 1957; Messick, 1989).

Developing a valid instrument for measuring a psychological construct is an iterative process of unpacking theoretical ambiguity into more tangible forms (Haucke et al., 2021). Depicting what content is deemed suitable to operationalize a construct is expressed by content domains that outline what attributes and characteristics are hypothesized to relate to the phenomena (Maul, 2017). Guidelines reiterate this clarifying step because it is more difficult to justify what content should be included when there are no specifications for what constitute valid metrics of the construct (Boateng et al., 2018; Clark & Watson, 2019; DeVellis, 2012). After all, how can scores from a test be interpreted as valid if it is unknown whether a test's items even represent a construct in the first place?

In this section, we articulate our content domains to operationalize all the Episode Model constructs. With the Episode Model integrating multiple explanatory approaches and conceptualizing a wide scope of constructs, it was necessary to develop multiple sets of items simultaneously. Here, we expand upon the model by defining specific attributes and characteristics of sub-constructs and offer a depiction of a hierarchical structure that may inform future measurement models.

Conceptual Basis – Episode Model

The Episode Model (Eerola et al., 2025) is a constructivist account of music–emotion induction that incorporates affectual self-regulation of emotional experiences to music as functional and situated emotional episodes. The theory describes its efforts as "… broadening the topic from stimulus-driven processes to goal-driven processes by allowing situation, motivation, and decision-making to be a substantial part of the theory" (p. 595, Eerola et al., 2025). Therefore, operationalizing this theory should include factors that facilitate the emergence of such emotional experiences.

Five emotional episodes were derived from empirical findings of common functions for music across different

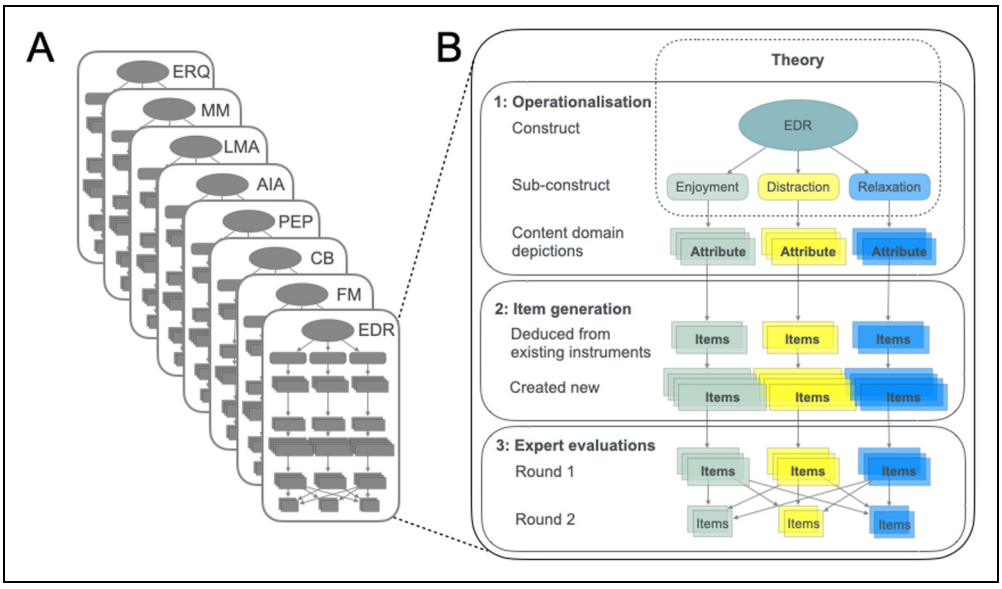

Figure 1. Outline of the scope to convert (A) eight constructs from the theory into separate sets of validated measurement items with detail of the processes with the example of (B) one construct proposed from the Episode Model.

cultures (Mehr et al., 2019) and situated uses of music (Randall & Rickard, 2017). These were consistent with previously conceived context-dependent affectual regulation (Baltazar & Saarikallio, 2016). These five episodes are as follows, (1) Enjoyment-Distraction-Relaxation (EDR) refers to "an affective process of enjoyable tension reduction" (p. 597); (2) Connection-Belonging (CB) was defined as "episode[s] in which the main functional focus is on social connection" (p. 597); (3) Focus Motivation (FM) conveys "experiences in which music supports task-related motivation and focus" (p. 597); (4) Personal-Emotional-Processing (PEP) was defined as "experiences of identifying with the musical content and allowing music to support emotional coping" (p. 598); (5) Aesthetic-Interest-Awe (AIA) was theoretically defined as episodes that pertain to "being moved, spirituality, detached emotions, aesthetic, or awe experiences" (p. 598). Each of these episodes types constitute an unobservable psychological process; thus they are considered henceforth as sub-constructs for the broader construct of Functional Context.

According to the Episode Model (Eerola et al., 2025), the five emotional episodes may be influenced by changes in different *descriptive schemes,* each of which is hypothesized to correlate with particular episodes to different degrees. Descriptive schemes refer to either psychological processes (emotion qualia, induction mechanisms, reward, and exposure), or important aspects such as attention, agency, or meaning known to contribute to emotional experiences (p. 599). Functional Context (FC) of the episodes relates to various functions of music, typically studied as emotion regulation goals. Listening Modes and Agency (LMA) consists of differences in attention given to music in a situation (Weining, 2022), and agency refers to ability to control or choose the music (Krause et al., 2015; Saarikallio et al., 2020). Musical Meanings (MM) denote (a) pathways of referential associations such as structure of music, (b) sensitivity to the source including the cultural context of music, and (c) personal significance and autobiographical memories of the piece of music (Thompson et al., 2023). Exposure, Reward, and Qualia (ERQ) relate to either stimulus exposure (e.g., preferences, familiarity; see Huron, 2006), or traditional emotion descriptors (core affects, Russell & Barrett, 1999 and music-related emotion qualia, e.g., AESTHEMOS; Schindler et al., 2017). Finally, one of the descriptive schemes represents the well-known emotion induction mechanisms (Juslin, 2013; Juslin & Västfjäll, 2008).

Expanding Sub-Constructs into Content Domain Depictions

Constructs proposed by the Episode Model are multifaceted, with several sub-constructs alluded to but not fully defined. Going beyond the model requires detailing these sub-constructs; for example, EDR episodes consist of three distinct areas – Enjoyment, Distraction, and Relaxation – each of which consist of their own unique identifiers. Across the 5 episodes and the 3 descriptive schemes investigated here, 25 of these sub-constructs were necessary to define and to expand the existing theory. Acknowledging the broader literature, we operationally defined what we considered suitable content for each sub-construct (Table 1). A hierarchical structure, from constructs to sub-constructs and subsequent items derived from content domains, provided an initial framework upon which to develop the instrument.

The 25 sub-constructs listed in Table 1 were hypothesized to cover all aspects necessary to investigate the

Table 1. Hierarchical abstraction, from broad constructs to content domains for each sub-construct.

Constructs	Sub- constructs	Content domains
Episodes		
Enjoyment-Distraction-Relaxation (EDR)	Enjoyment	Music used for the positive experience of itself or to conjure desired effects from entertainment, dance, play
	Distraction	Music used to get away from worries and stressors, mediate boredom
	Relaxation	Music used to calm down, decompress, achieve low arousal state
Connection-Belonging (CB)	Group cohesion / socialization	Music used to facilitate connection, bonding, express joint identity or values, show commitment to a group
	Reducing loneliness	Music used to mediate loneliness, become a part of something, as a tool to dispel feelings of isolation
Focus-Motivation (FM)	Energy control	Music used to control levels of intensity or drive, pumping up, motivation, stimulation
	Focus	Music used to maintain or enhance a mood, organize self or group coordination to a task
Personal Emotional Processing (PEP)	Reflection / coping	Music used to remember or to associate to other objects, provide sense of comfort, processing experiences
	Expressing feelings	Music used to release or communicate emotions, gain emotional insight, deal with a situation
Aesthetics-Interest-Awe (AIA)	Being moved / spirituality	Music used to connect to a higher ideal or abstract feeling of non-social connection, religious activity, feeling or seeking strong emotions
	Curiosity	Music used to prompt new experiences, drive or attraction toward stimuli
	Aesthetics	Music used to achieve or create a desired ambience, fit to the "mood"
Descriptive schemes		
Listening modes & agency (LMA)	Diffuse	Little to no attention directed at the music, choice of focusing on objects other than music
	Bodily	Potential embodied motor processes are attuned to the music, agency (or lack thereof) over body response toward music
	Emotional	Reflects attention directed toward the intensity of emotional experiences one could have with music, level of control over emotions experienced
	Associative	Ideas, thoughts, and images prompted by the music, control over music to guide experience
	Structural	Listening emphasis on processing musical content; understanding structure, themes, syntax, or patterns of sound.
	Reduced / causal	Careful, forensic analysis of sound done with a reflective mindset
Musical meanings (MM)	Musical structure	Imbued meaning associated to musical elements that are consciously important
	Self	Personal memories (couples songs, associations to people), self-expression
	Source	Sociopolitical context, historical conditions, musicians' personae, performative interpretations
Exposure, reward, & qualia (ERQ)	Exposure	Mere exposure to stimuli, perception of stimuli change, reward of finding something new
	Familiarity	Recollection of prior instances of liking stimuli, grouping of similarity, classification reward
	Preferences	Conscious liking toward stimuli, account of personal value, value judgements, rewarding interactions
	Physical reactions / qualia	Conscious acknowledgement of physiological sensations, or physical pleasure (sensory pleasure/displeasure)

predictions as proposed by the Episode Model. Each sub-construct comprises its own body of research, so further review of the literature may uncover additional aspects of each concept to incorporate, covered in the next section. While these content domains are limited by the bias of the test developers, they cover the known conceptual basis outlined by the Episode Model and will guide generation of a relevant item sample.

Translating Content Domains into Items

Building an item sample requires specification of attributes or characteristics assumed to reflect a construct, so that all

Table 2. List of existing instruments utilized in deductive item generation with IDs for subsequent use.

ID	Domain of interest	Name of questionnaire	# of items adapted
U	Functional musical engagement	Uses of Music Inventory (Chamorro-Premuzic & Furnham, 2007)	14
F	Functional musical engagement	Functions of Music Listening (Schäfer et al., 2013)	115
A	Attentional concepts	Mindful Attention Awareness Scale (Brown & Ryan, 2003)	3
M	Attentional concepts	Cognitive and Affective Mindfulness Scale (Feldman et al., 2007)	5
K	Attentional concepts	Kentucky Inventory of Mindfulness (Baer et al., 2004)	8
R	Emotion regulation	Music in Mood Regulation (Saarikallio, 2008)	39
G	Emotion regulation	Goldsmiths Music Sophistication Index[1] (Müllensiefen et al., 2014)	13
E	Emotion regulation	Emotional Regulation Questionnaire (Gross & John, 2003)	4
C	Agency with music	Desire for Control Over Listening Scale (Krause et al., 2020)	20
B	Musical reward	Barcelona Music Reward Questionnaire (Mas-Herrero et al., 2013)	17
S	Aesthetic concepts	Aesthetic Experience Questionnaire (Wanzer et al., 2020)	21
P	Cog/Social aspects of music listening	Music Empathizing/Systemizing Inventory (Kreutz et al., 2008)	14
O	Cog/Social aspects of music listening	Absorption in Music Scale (Sandstrom & Russo, 2013)	17

1Emotions subscale.

relevant content can be incorporated from the universe of possible items (Boateng et al., 2018; Clark & Watson, 2019). Following this best practice, we sought to create an initial item sample that would address all known aspects of each sub-construct outlined in Section 1. Starting from a literature review of related measurement tools, a deductive method was used to gather relevant items from pre-existing instruments (Table 2). A deductive method identifies a relevant item sample through logic, derived from literature concerning the construct and tangentially related concepts (Kline, 2000). Categorization of items into specified content domains required organizing the theoretical space into a matrix (Table 3). This matrix differentiated the content domain of each sub-construct through a framework for perceptual and cognitive evaluation of environmental change – with cognitive, affective, motivational, and social variables (Rauthmann, 2015). This consideration reframed the content domain for each sub-construct, positioning the original content domains for the purpose of systematic content generation but also offering deeper theoretical distinction (Table 1). After deductive generation from existing instruments, this matrix highlighted underrepresented content areas. New items were created from concepts in the literature to fill these gaps. All the items were adapted to follow a reflective writing style so the instrument could be administered to capture ongoing or recently experienced emotional episodes to music. For example, an item from the Barcelona Music Reward Questionnaire (BMRQ; Mas-Herrero et al., 2013), "Music calms and relaxes me," was changed to "The music made me feel calm or relaxed." Finally, we assessed the semantic similarity of this initial sample to examine high-similarity items for rewriting or removal.

Deductive Item Generation

We (CK, TE) started by reviewing existing instruments, both within and outside music psychology literature, to find existing items to use as indicators for our sub-constructs. Thirteen pre-existing instruments were identified (Table 2). Some content areas were left absent during item generation (see Table 3, e.g., see row "Physical reactions / Qualia" and column "Motivational") because well-established constructs (i.e., valence and arousal) cover these areas (labelled "V&A" in Table 3). In total, 13 pre-existing instruments were identified as having relevant content for the sub-constructs (290 items), although not all the items from each instrument were used. A matrix visualized the extent to which pre-existing items informed content domains (Table 3), and we found that the sample of items collected from the deductive generation did not fully operationalize the Episode Model. Therefore, new items were created to fill matrix gaps (205 items). For instance, "This was the music people said I needed to use in order to focus" (see row FM – Focus and column Social in Table 3) was created by reviewing literature on social influences of music preferences and performing cognitive tasks alongside music (e.g., Bonneville-Roussy et al., 2017; Vigl et al., 2023).

Item Curation

Following the generation of initial content, the item sample consisted of a mixture of grammatical styles and response formats; hence we had to edit some of the items and to remove problematic ones (e.g., double-barreled items, "not" items, items that may be universally endorsed or rejected, and those containing expressions or colloquialisms; see Fowler, 1995). Pre-existing items were reworded to position responses to reflect recent states (e.g., "I felt…" rather than "I typically feel…"). Items generally specified or targeted music as the object addressed in the situation or experience (see examples in Supporting Information, Eerola & Kirts, 2025). Additionally, each item's meaning was investigated by experts at a later stage (Section 3) to discover whether items effectively communicated the intended meaning assumed after generation according to the content domain representing the sub-constructs.

Table 3. Counts of items across episode constructs and situational variables.

Constructs	Sub- constructs	Situational variables				Pre-existing	New
		Cognitive	Affective	Motivational	Social		
Enjoyment Distraction Relaxation (EDR)	Enjoyment	F^3, B, R, N	N^5	F^4, N	C^3, N^4	12	11
	Distraction	M, F^2, K, O^2	R^4, N	K, O, F^4	F, N^5	17	6
	Relaxation	B^2, R, F, S, N^2	R^2, N^4	F^2, N^4	N^7	9	17
Connection Belonging (CB)	Group cohesion / socialization	O^2, B^2, F^2	F^2, U, N^2	F^4, N	F^2, C^4, N^2	19	5
	Reducing loneliness	F^3, B^2, O, N	F, R, N^2	F^4, U, N	C^2, R, B, N^3	17	7
Focus Motivation (FM)	Energy control	F^2, N^4	R^4, F^2, U	F^4, N^2	C, N^4	14	10
	Focus	M^3, U, O, N^2	R, U, G, N^2	F^6	N^5	14	9
Personal Emotional Processing (PEP)	Reflection / coping	R^2, F, B, N^2	R^7, F	F^4, N^2	N^6	16	10
	Expressing feelings	F^3, R, N^2	R^4, F^2, E	F^2, N^3	E, R, N^4	15	9
Aesthetics Interest Awe (AIA)	Being moved / spirituality	O^2, F, S, N^2	R, N^3	F^2, O, N^2	N^5	8	12
	Curiosity	F, S, R, U	N^4	F, N^4	C, N^4	6	12
	Aesthetics	F^2, R^2, N^2	R, N^3	N^5	N^4	5	14
Listening Modes & Agency (LMA)	Diffuse	A, N^4	A, N^3	A, F, N^2	C^2, N^3	6	12
	Bodily	B^2, F, R, S, N	F, N^4	F^5, O	N^5	12	10
	Emotional	B^2, U^2, K, S, N	G^2, U, S, N	N^4	P^2, E, N^2	13	8
	Associative	F^3, O^2, K, R, U	G, K, N^3	F^5, G, S	F^2, E, N^2	20	5
	Structural	K, S, P, O, N	P, N^3	F, N^3	G^4, N^2	10	9
	Reduced & causal	S^2, K, M	U, N^2	U, N^3	S, N^5	7	10
Musical Meaning (MM)	Musical structure	S^2, P, N^3	P, N^4	F^3, N^2	S^3, G, N^3	11	12
	Self	F^4, P^2, O	U, R, N^3	F^5	F^3, C^2, N	19	4
	Source	F^3, S^3, O^2, P	U, P, N^3	F^5	P^3, S^2, N^2	21	5
Exposure, Reward, & Qualia (ERQ)	Physical reactions / qualia	F^2, B, R, K, P	B	V&A	G	8	0
	Preferences	B^2, G, N	V&A	F^2	C^3	8	1
	Familiarity	G, N^5	V&A	N	C^2, N	3	7
Pre-existing		112	54	73	51	290	
New		34	52	40	79		205

Note. The items are depicted by their source (See Table 2 for the instrument IDs) with item counts (e.g., N is a single item, whereas N^5 has five items from the same source). New items are indicated with "N," while "V&A" refers to valence and arousal.

Results – Initial Item Sample

In total, 495 items were adapted or created through the deductive item generation. Of these, 290 items were found in 13 pre-existing instruments and 205 items were created specifically for this instrument. In Table 3, counts of the items across constructs and sub-constructs are shown in superscript and capital letters indicate origin of the item. For the EDR construct, the Enjoyment sub-construct is represented by six items (indicated as F^3, B, R, N). Three items were obtained from the Functions of Music Listening survey (abbreviated F; Schäfer et al., 2013), one from the Barcelona Music Reward Questionnaire (B; Mas-Herrero et al., 2013), one from Music in Mood Regulation (R; Saarikallio, 2008), and one bespoke item (N). Some instruments feature prominently due to conceptual overlap and because they are extensive inventories (e.g., Schäfer et al., 2013 with 115 items). The Supporting Information reports all questionnaire items (see Eerola & Kirts, 2025). As a starting point, this collection of items was assumed to be comprehensive but also excessive for any practical use.

Reduction of the Item Sample. To make later content validity examinations feasible for subject matter experts (SMEs), we (TE, CK) subjected the 495 items to semantic similarity analysis to eliminate close variants of the same question coming from different sources (Figure 2 provides a comprehensive overview of the full reduction of the item pool). Semantic similarity analysis was conducted by calculating cosine similarities between every item using word vector embeddings included in the *spaCy* natural language processing (NLP) library (Honnibal et al., 2023), relying on 683,000 unique vectors in 300 dimensions (*en_core_web_lg*

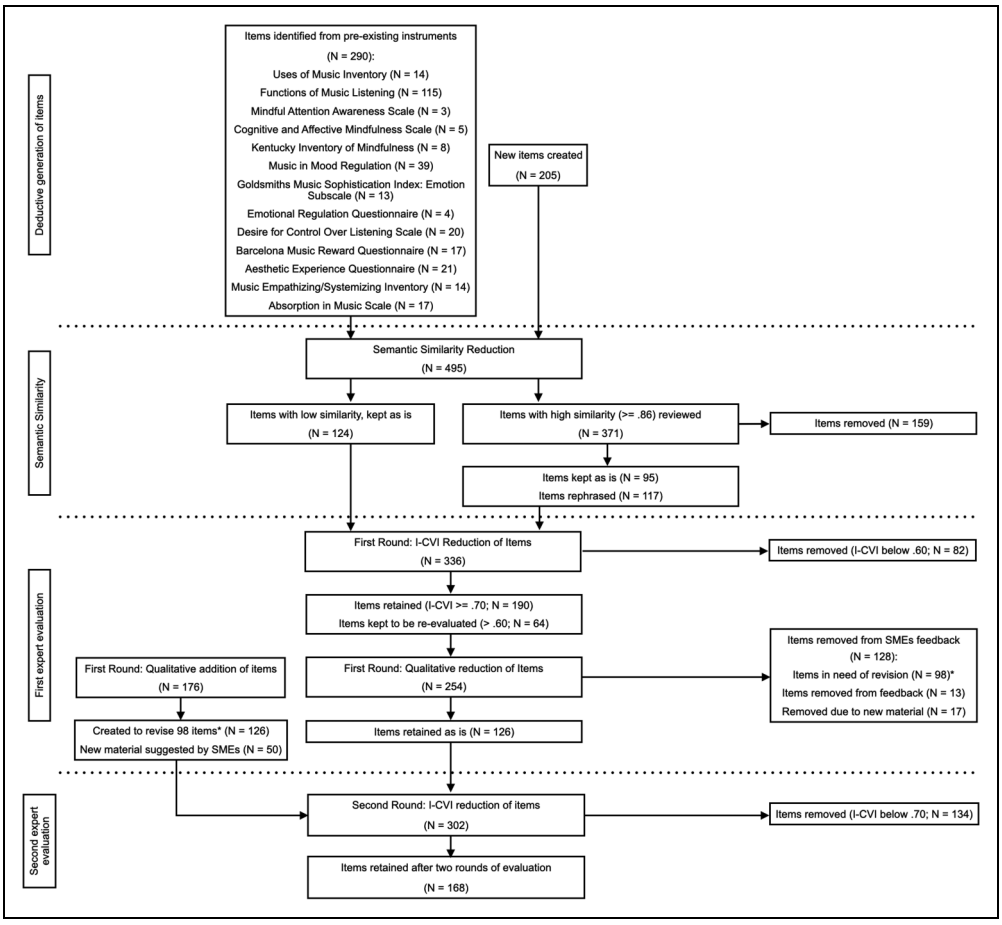

Figure 2. Flowchart detailing the reduction and addition of items. From the initial pool of 495 items generated to the second round of evaluation by SMEs.

version 3.8.0). Items with high similarities (>= .86 on a scale of 0–1; e.g., "I liked the music" and "I liked hearing music") were manually checked ($N = 371$) by CK so they could either be rephrased ($N = 117$), kept as is for further evaluation ($N = 95$), or removed ($N = 159$). This operation resulted in a pool of 336 items, making further evaluation a more feasible process for SMEs. However, it remained uncertain whether the 336 items adequately reflected the intended constructs. Likewise, it was unclear if the content domain had been comprehensively sampled. To explore these issues, we proceeded with an empirical investigation of the item sample to assess relevance and representativeness of the items and to identify any conceptual gaps.

Evaluating Content Validity

While a deductive method of item generation should capture all relevant content, bias of the test developers may sway the resulting item pool if hypotheses concerning item relevance are not tested. Hence, proper item generation also needs an inductive method to generate content missing from the initial sample and revise items from insight and experience, usually with SMEs (Boateng et al., 2018; Clark & Watson, 2019). There is an unresolved discussion about the usability of results from these evaluations (see Beck, 2020). However, it is generally agreed that an expert panel can (1) identify missing content, (2) indicate clarity and understandability, (3) judge content relevance, and (4) identify representative items (Almanasreh et al., 2019). Furthermore, collaboration with other scholars has the added benefit of incorporating multiple opinions and voices from varying positions on the topic. This ensures that if there were major gaps in the operationalization from the theory, or the theory itself, experts would highlight these issues. As the final step for the first stage of development, the process of evaluating content validity is critically important for the development of any psychometric instrument (Sireci, 1998).

Our inductive investigation was conducted in two rounds, administered online through Qualtrics. This process could continue indefinitely, but two rounds provide a quick initial development, with the first incorporating expert insight and the second re-evaluating this feedback and solidifying a highly relevant and representative item pool. The SMEs were asked to provide insight about which items they considered relevant, say whether

content was under-represented, and describe any conceptual gaps in the items or model. Each item was assessed by quantitative ratings of relevance (Rounds 1 and 2), representativeness ratings (Round 2), and optional qualitative responses (Round 1 and 2). The SMEs were given resource documents describing their involvement, ethical information, the purpose and aim, and definitions of constructs and sub-constructs (see Supporting Information, Eerola & Kirts, 2025).

First Round: Procedure

Eight surveys, 1 for each construct, were administered to assess the 336 items. The items were randomized within each survey, and participants followed different survey orders to reduce ordering effects. Fourteen SMEs participated in the first round: eight based in Europe, two in South America, two in North America, one in Asia, and one in Australia. Two were in the process of obtaining a PhD, six were early career researchers, and six were established faculty or industry professionals. All participants had engaged with music psychology, either through their education, research, or teaching. Eight were native English speakers (US, Canada, and UK), and the rest (7) had published multiple articles in English and operated successfully in an English-speaking academic context (currently working in an English-speaking context or having done a degree in the UK or US). Seven of the experts had previous experience of scale development. The SMEs provided consent to participate before starting each survey. Items were displayed one at a time with two relevance questions. The first asked how relevant they thought the item was on a four-point scale (1 = *not relevant* to 4 = *extremely relevant*). The second was an optional open response for SME feedback (e.g., face validity, importance, rejection rationale, corrections). At the end of each survey, the SMEs could provide feedback concerning the construct's devised content (e.g., challenging definitions, relations to other variables, missing concepts).

First Round: Item Reduction and Interim Results. Reduction after the first round involved both quantitative and qualitative analyses. Quantitative ratings were analyzed with item-level indications of content relevance (I-CVI), which measure agreement by calculating the proportion of experts rating an item as being relevant on a four-point scale (Eerola, 2024). As a practical example, an I-CVI of 85% would mean that 11 out of 13 experts agree that an item is "highly relevant" for the construct. Using an established I-CVI cut-off score of .70 or above (Polit & Beck, 2006), we retained 190 items, cut 82 items (below .60) and kept 64 items slightly below the cut-off to be reevaluated in the second round.

Qualitative expert feedback was used to improve the 254 items remaining after quantitative analysis. Ninety-eight items needed rewriting to clarify the intended meaning or to separate aspects of the underlying content (creating 126 items; see Supporting Information: Eerola & Kirts, 2025), 13 items were removed based on overwhelming feedback, and 17 items were removed because better items of the same meaning proposed by the SMEs made them redundant. For example, experts found that many items referencing religious experiences were problematic, prompting refinement. They also raised concerns about statements relating to metaphysical concepts of "Soul," which prompted removal. For example, SMEs pointed out the attribution of "achiev[ing a] low arousal state" to EDR's Relaxation subconstruct would be more aptly conveyed as "reducing tension." In total, 126 of the 254 items were retained without revision, and an additional 176 items were created based on SMEs feedback. Of the additional items, 126 were rephrased from the 98 items identified as having issues (29 of those being made to ensure theoretical coverage) and 50 items were suggestions of missing content proposed by the SMEs. In total, 302 items continued to the second content evaluation.

Second Round: Procedure

For the second round, all previous participants were asked to re-evaluate a reduced item pool following their feedback from the first round. Twelve SMEs returned and one new senior researcher joined ($N = 13$): eight were based in Europe, two in North America, one in South America, one in Asia, and one in Australia. Three surveys were distributed for the second round – one for all episode items; one for Listening Mode and Agency (LMA) items; and one combined Musical Meaning (MM) and Exposure, Reward, and Qualia (ERQ) items. The SMEs rated each item with the same four-point relevance scale, but items for a sub-construct were shown together alongside the content domain depiction. Grouping the content in this way allowed the SMEs to review whether items discriminated between the sub-constructs. The SMEs also ranked a sample of items they considered to be the most *representative* of the sub-construct. Finally, the SMEs could express concerns with open response questions for each sub-construct and at the end of each survey. The second evaluation used the same reduction methods as the first (I-CVI for each item and feedback) alongside representative item nominations to evaluate what aspects of content domains might be lost.

Main Results Across the First and Second Evaluation Rounds

To summarize ratings, item level values (I-CVI) from both rounds were collapsed into scale level values of content relevance (S-CVI, see Romero Jeldres et al., 2023) that report average relevance for each item pool (N). Round 1 elimination from expert relevance agreement (I-CVI scores below .70, see Polit & Beck, 2006) reduced the initial item pool to 190 items, with an average S-CVI of .93 across all sub-constructs. Re-evaluation from Round 2 reduced the item pool to 168 items, with a .92 average S-CVI across sub-

Table 4. Descriptive progression of SMEs rating content as relevant (S-CVI) and representative.

		Round 1 (SMEs = 14)		Round 2 (SMEs = 13)		
Construct	Sub-construct	S-CVI: All items	S-CVI: Top items R1	S-CVI: Top items R2	Top 5 most representative items ID (agreement)	Construct S-CVI
EDR	Enjoyment	.69 (N = 13)	.89 (N = 6)	.87 (N = 7)	21, 16, 18, 19, 22 (77%)	.90 (N = 22)
	Distraction	.72 (N = 17)	.89 (N = 7)	.91 (N = 8)	27, 36, 29, 28, 33 (81%)	
	Relaxation	.88 (N = 16)	.94 (N = 11)	.90 (N = 7)	1, 9, 10, 6, 8 (83%)	
CB	Group cohesion / socialization	.82 (N = 17)	.96 (N = 9)	.94 (N = 9)	1, 11, 2, 10, 13 (81%)	.94 (N = 15)
	Reducing loneliness	.80 (N = 15)	.93 (N = 10)	.95 (N = 6)	28, 19, 24, 21, 20 (78%)	
FM	Energy control	.62 (N = 17)	.90 (N = 5)	.95 (N = 3)	8, 2, 4, 10, 3 (71%)	.94 (N = 14)
	Focus	.74 (N = 14)	.98 (N = 4)	.94 (N = 11)	11, 25, 13, 14, 18 (84%)	
PEP	Reflection / coping	.86 (N = 14)	.98 (N = 9)	.95 (N = 10)	2, 6, 9, 11, 12 (87%)	.95 (N = 20)
	Expressing feelings	.80 (N = 16)	.94 (N = 9)	.95 (N = 10)	13, 14, 15, 23, 24 (78%)	
AIA	Being moved / spirituality	.81 (N = 12)	.95 (N = 7)	.90 (N = 6)	6, 2, 1, 8, 11 (86%)	.90 (N = 11)
	Curiosity	.73 (N = 15)	.89 (N = 8)	.88 (N = 2)	14, 15, 13, 19, 21 (76%)	
	Aesthetics	.75 (N = 14)	.94 (N = 6)	.92 (N = 3)	26, 27, 29, 34, 28 (82%)	
Listening Modes & Agency	Diffuse	.76 (N = 14)	.93 (N = 5)	.96 (N = 9)	4, 2, 12, 3, 9 (88%)	.94 (N = 53)
	Bodily	.79 (N = 18)	.93 (N = 10)	.93 (N = 8)	3, 13, 1, 2, 4 (87%)	
	Emotional	.80 (N = 13)	.90 (N = 8)	.91 (N = 16)	2, 15, 7, 1, 10 (85%)	
	Associative	.82 (N = 18)	.93 (N = 12)	.94 (N = 7)	4, 12, 1, 3, 7 (89%)	
	Structural	.83 (N = 8)	.94 (N = 6)	.96 (N = 8)	2, 5, 9, 4, 1 (88%)	
	Reduced / causal	.86 (N = 11)	.98 (N = 8)	.93 (N = 5)	1, 3, 6, 7, 10 (87%)	
Musical Meanings	Musical structure	.77 (N = 16)	.89 (N = 7)	.91 (N = 4)	2, 5, 9, 4, 1 (88%)	.91 (N = 21)
	Self	.86 (N = 19)	.95 (N = 14)	.92 (N = 10)	5, 1, 11, 3, 4 (87%)	
	Source	.90 (N = 19)	.96 (N = 15)	.89 (N = 7)	8, 15, 1, 13, 7 (89%)	
Exposure Reward, & Qualia	Familiarity	.90 (N = 6)	.95 (N = 5)	.92 (N = 6)	13, 2, 9, 5, 10 (86%)	.95 (N = 12)
	Preferences	.66 (N = 8)	.95 (N = 3)			
	Physical reactions/qualia	.96 (N = 6)	.96 (N = 6)	.97 (N = 6)	1, 6, 5, 4, 2 (86%)	
Total:		336 items	190 items	168 items	115 items	168 items

Note. *N* refers to the number of items in the construct, *most representative* refers to the item number (#) nominated by most experts as the most representative item, and % indicates overall agreement of these nominations.

constructs. Additionally, expert agreement about which items were the most representative, referred to by ID, is reported in Tables 4 and 5. Finally, Table 3 reports construct S-CVI after the second round of evaluation (all item I-CVI values for both rounds are available in the Supporting Information, Eerola & Kirts, 2025).

From Round 1, consistently high values of agreement – ranging between .62 (Energy Control: FM) to .96 (Physical reactions: ERQ) – were yielded for all the items in each sub-construct (Table 4; column "S-CVI Initial" – this represents the content validity index from SMEs). All sub-constructs received values above the .60 S-CVI threshold for 14 experts (Exact method; see Romero Jeldres et al., 2023). The most consistent items for each sub-construct (190 items total), scoring above .70 I-CVI, are shown under S-CVI Top Items where values range from .89 to .98.

From Round 2, items which scored above .70 I-CVI (S-CVI sub-constructs, Table 4) left 168 items, with means of 21 per construct and 7 per sub-construct, ranging between 2 (Curiosity) and 16 items (Emotional). Representative agreement for the top five most nominated items ranged from 71% to 89%. Finally, the last column of Table 4 shows high relevance agreement for each construct (S-CVI across sub-constructs), ranging from .90 to .95 (with all top scoring sub-construct items, N = 168). As S-CVI values indicate the proportion of all experts giving "quite relevant" or "very relevant" scores for the items, consistent values above acceptable S-CVI thresholds of 0.80 (Polit & Beck, 2006) or even 0.90 (Waltz et al., 2005) is reassuring. To examine the contents of the sub-constructs in more detail, we summarize the two most representative items from this assessment in Table 5. For a more extensive list, see the Supporting Information, Eerola & Kirts, 2025.

Table 5 presents what our expert panel considered the two most representative items for each sub-construct. Enjoyment, for example, contains items like "Listening to music was a good way of entertaining myself" (EJ13, I-CVI = 0.92; 47% expert agreement) and "The music gave me pleasure" (EJ2, I-CVI = 0.92; 31% agreement). It is worth noting that, in most cases, the two most representative items were unsurprisingly deemed relevant (I-CVI scores typically above 0.90, with one exception – 0.62 – for the second most representative item under Connection). As the representativeness index reflects proportional agreement for items nominated by the SMEs, the top two items ranged from 100% to 31% agreement. It remains to be seen how items operate in actual use, but theoretically consistent observable indicators for each sub-construct have emerged through this empirical process, improved further by valuable opinions from our expert panel. Overall, the insights from qualitative comments from the SMEs suggested that while it was a challenging task at times, the multiple iterations helped them to articulate what the sub-constructs are through selection and refinement of the agreed items.

General Discussion and Conclusion

Our primary aim was to formulate and implement operational definitions of the constructs outlined by the Episode Model into a psychometric instrument (Eerola et al., 2025). Here, we documented our process of creating and evaluating content made to represent these constructs. This includes (1) our interpretation of the Episode Model's theoretical constructs into operationalized content domains depicting attributes and characteristics for 25 sub-constructs; (2) the process of generating items that reflect the content domains for each sub-construct, by adapting pre-existing items or creating new ones; and (3) our empirical investigation of this content with subject experts who reviewed theoretical positioning and items' coherence to content domains, and suggested missing content to improve the item pool. From our articulation of the constructs and demonstration of what constitutes suitable content to reflect these constructs, we hope that researchers will be able to more clearly and easily assess whether this instrument is suitable for their purposes. The development of the sub-constructs as independent modules also provides researchers with a conceptual starting point for the development of new instruments, using content outlined here as a foundation.

Our secondary aim was to gain insight into existing items for constructs commonly used in music psychology research. A deductive item generation method connected the present work to existing and tangential conceptualizations of each construct and adopted items from pertinent existing instruments. Just over half (58.6%) of the items initially generated originated from existing instruments. Content derived from these pre-existing instruments typically contributed to 3–5 constructs under this model. However, items from a given instrument did not always end up within the same construct or even in expected constructs given the existing instrument's reported purpose. As the present work does not assess the success of existing instruments' ability to capture their intended constructs and since their items were rephrased here to ensure consistency in the wording for our purpose, this insight is not presented as a critique but as noteworthy consideration for future studies.

In the final section, SMEs examined whether this item sample was appropriate for our sub-constructs. As a consequence of involving colleagues from around the world, cultural and English-speaking colloquialisms were phased out and more succinct articulations of the intended message/meaning of items were offered to improve the item sample. A first evaluation established relevant content for all the sub-constructs, while feedback added missing content and clarified language. Expert feedback improved the initial item sample by identifying the most suitable items for each sub-construct and construct beyond the bias and knowledge of the test developers. Confirmation of relevance carried into the second evaluation where the SMEs also indicated highly representative items for each sub-construct, making it possible to review which items

Table 5. Two most representative items from each sub-construct.

Construct	Sub-construct	Item (repository ID [Table 4 ID], I-CVI, representativeness agreement)
EDR	Enjoyment	Listening to music was a good way of entertaining myself (EJ13 (21), 0.92, 46%)
		The music gave me pleasure (EJ2 (16), 0.92, 31%)
	Distraction	Music acted as a mental shield against my distressing thoughts (DT7 (27), 0.92, 54%)
		The music pushed my worries aside (CC5 (36), 0.92, 46%)
	Relaxation	The music made me feel calm or relaxed (RX1 (1), 1.00, 69%)
		I used the music to soothe my mind (RX57 (9), 0.92, 54%)
CB	Group cohesion / socialization	I felt more connected with other people because of the music (GC1 (1), 1.00, 69%)
		I thought of sharing this music with my peers (GC54 (11), 0.62, 38%)
	Reducing loneliness	The music made me feel like I belonged (SL53 (28), 0.92, 46%)
		I used the music to keep me company (SL57 (19), 0.77, 38%)
FM	Energy control	Listening to music made me more alert (EM14 (8), 0.92, 46%)
		The music was motivating me to do something (EM50 (2), 1.00, 31%)
	Focus	Music helped me focus on what I was doing (MF2 (11), 1.00, 69%)
		I used the music to motivate myself (MF59 (25), 1.00, 54%)
PEP	Reflection / coping	The music helped me process my emotions (CC3 (2), 0.92, 100%)
		I used the music to cope with my situation (CC17 (6), 1.00, 62%)
	Expressing feelings	I vented my emotions by listening to music that expressed my feelings (EF6 (13), 1.00, 46%)
		Listening to music helped me express how I was feeling (EF12 (14), 1.00, 38%)
AIA	Being moved / spirituality	The music provided me with a spiritual experience (MS53 (6), 0.92, 85%)
		I was completely immersed in the music, as it altered my state of being (MS50 (2), 1.00, 62%)
	Curiosity	The music sparked my curiosity (CU13 (14), 0.85, 46%)
		Listening to the music was an intellectual experience for me (CU4 (15), 0.77, 46%)
	Aesthetics	I appreciated the beauty of the music (AE3 (26), 1.00, 69%)
		The experience was enhanced by the beauty of the music (AE50 (27), 0.85, 46%)
LMA	Diffuse	Music was simply in the background for me (DF52 (4), 1.00, 85%)
		I was more attentive to what I was doing than to the music (DF50 (2), 1.00, 69%)
	Bodily	It felt natural to tap or move along with the music (BA51 (3), 1.00, 69%)
		I wanted to move along with the music (BA56 (13), 0.83, 62%)
	Emotional	My focus was on my emotions while I was listening to the music (EA16 (2), 1.00, 69%)
		I was in control of how the music made me feel (EA57 (15), 0.92, 54%)
	Associative	I realized the music caused visual imagery in my mind (AA53 (4), 1.00, 85%)
		I realized I was creating a fictional story or scene along with the music (AA54 (12), 0.92, 77%)
	Structural	I chose to pay attention to how the music was developing (SA51 (2), 0.92, 85%)
		I focused on hearing patterns in the music (SA53 (5), 1.00, 62%)
	Reduced	I gained new insights about the music by focusing on the small details (RC2 (1), 0.92, 77%)
		I was analyzing the complexity of the music (RC51 (3), 1.00, 62%)
MM	Musical structure	How the music was structured is important to me (ST22 (2), 0.93, 77%)
		The harmony, melody, or rhythm prepared me for upcoming changes (ST13 (5), 1.00, 69%)
	Self	The music reminded me of a certain time in my life (SE54 (5), 1.00, 69%)
		I found the music to be personally meaningful (SE50 (1), 0.92, 62%)
	Source	I considered the music's historical context to understand what it meant (SO56 (8), 0.92, 92%)
		I connected to the message the artist(s) were trying to convey with the music (SO62 (15), 1.00, 85%)
ERQ	Physical reactions / qualia	I experienced physical sensations (tears, shivers, goosebumps) while listening to the music (QU50 (1), 1.00, 69%)
		While listening, I noticed tears starting to form in my eyes (QU53 (6), 1.00, 69%)
	Preference & familiarity	Listening to the music was a rewarding experience (CU2 (13), 0.92, 77%)
		I knew I had never heard the music before (PO50 (2), 0.77, 54%)

Note. The values in brackets refer to the unique item identifier, I-CVI score for the item, and proportion of experts nominating the item to be a representative. LMA refers to Listening Modes & Agency, MM refers to Musical Meaning, and ERQ refers to Exposure, Reward, & Qualia.

are necessary from the relevant sample. The descriptive data regarding item quality (relevance and representativeness scores) can be used in the future to gauge suitability for later test development processes. For example, some sub-constructs may not have sufficient items for some latent factor processes, although "lower" quality items could be used considering their moderate I-CVI values. At this stage, it remains unknown whether these items yield valid responses for a particular purpose and, while expert agreement constitutes an informed foundation, whether the item samples still reflect the content domain descriptions articulated here necessitates further empirical investigation. This future work should reevaluate the content validity of the item sample to ensure it continues to reflect intended constructs and, where necessary, update the content. Moreover, separate testing of hypotheses concerned with the internal consistency of this instrument will need to be conducted. For example, future research should investigate whether (1) the dimensionality for each construct matches the assumed hierarchical structure presented here, (2) how response behavior changes the internal structure of sub-constructs, and (3) differences between populations indicate similar internal structure for the constructs. Additionally, these empirical investigations should be used to further understand if the item sample reflects the content domains as they are defined here or if expert evaluation tilted the conceptual frame to reflect slightly different concepts.

Overall, our processes adhered to current test development standards regarding content validity (Boateng et al., 2018; Clark & Watson, 2019), with the expert evaluations filtering the item pool to 168 highly relevant items across 23 sub-constructs. While this can be seen as a positive outcome and part of the process of identifying suitable items, critical discussion regarding the scope of the underlying theory and its implications for measurement and theory testing is needed.

Scope and Modular Theory Design

It is worth discussing the broad scope of the theory, which necessitated developing constructs for several distinct domains of interest – ranging from listening modes and musical meaning to core affect. This posed a major challenge for scale construction as it effectively meant that we were developing a collection of related instruments rather than a single focused and unified measure of one target attribute. However, with a lack of validated self-report instruments for our constructs, we embraced the challenge of addressing this breadth. It was therefore necessary to explicate and test multidimensionality, not only for our theoretical reasons but because it is recommended even with highly homogeneous instruments, before inferring a validity interpretation due to the assertion that any single item contains multitudes of meaning and reflects numerous constructs (Borsboom, 2023; Clark & Watson, 2019; AERA et al., 2014, p.11). In this study, we acknowledge that alternative items could have captured the underlying theory equally well. What is crucial, however, is recognition of the valuable process of developing items through rigorous procedures to reflect a particular theory and offering a basis upon which validity inferences can be made.

Here, eight constructs were developed to serve a wider purpose of contextualizing emotional episodes. Additionally, each construct was operationalized with sub-constructs to explain divergent indications of the broader phenomena. We propose that each construct can be used as an independent module to probe specific aspects of emotional experiences related to music. Some of these are more directly aligned with the core concept of episodes, while others are more peripheral to contextualize each experience. This reflects the theory itself: Episodes are central, whereas descriptive schemes are optional and serve as additional ways of characterizing episodes (Eerola et al., 2025). For example, research focused on music and sports performance might only wish to concentrate on the *Focus-Motivation* construct (with its two sub-constructs, Energy Control and Focus) and perhaps add just two relevant descriptive schemes (*Preference and Familiarity*, and *Musical Meanings*) with the corresponding items, rather than employing the entire arsenal of items from all constructs. Such modularity provides flexibility for this item set to enhance applicability across different research contexts. In some cases, the 5 constructs corresponding to the 5 episodes would be the most relevant modules, currently consisting of 82 items, but this could be assessed with 10 to 25 items if only the 2 or 5 most representative items per construct are used. Other constructs – such as those concerning *Musical Meaning*, *Listening Modes,* and so on – can also be treated as independent modules, which may be useful for certain studies but not others. It may even be possible to select and bolster sub-constructs for a very specific focus of a study. For example, a study investigating the role of social context in musical experiences could fully employ *Connection-Belonging* items as a module while using a subset of items from other modules. Thus, modularity facilitates theory-driven research without burdening participants with unnecessarily broad instruments, depending on the aims of each study. However, it remains to be seen what the final measurement models are for the constructs after psychometric investigation.

Limitations and Future Research

While the breadth of the theory allowed us to formulate an initial measurement structure that is flexible enough to enhance applicability for different research contexts, optimization of items within each construct still requires psychometric investigation of the instrument. Looking ahead, we will apply the items in a new empirical evaluation involving non-expert participants to identify a core set of items and explore interrelationships with factor analysis. We argue that developing a flexible and context-sensitive instrument, rather than aiming for a universal theory with

an all-encompassing item set, serves a pragmatic and productive approach to advancing multiple lines of inquiry, including a procedural approach to music and emotion. Each module in our instrument may be an imperfect representation of its target construct, but each contains content assumed to be capable for their intended purpose. With this starting point, future researchers may wish to extend the work done here to refine the sample of items in light of new theory and empirical need.

Our review of the literature signaled that it is not common practice to thoroughly validate instruments – from strong theoretical predictions to hypothesize and test what content reflects constructs – although some examples have incorporated multiple forms of validity evidence (e.g., Groarke & Hogan, 2018; Henry et al., 2024; Krause et al., 2020; Law & Zentner, 2012). Theories concerning music and emotion, for example, have not benefited from the rigorous and systematic established standards of psychometric development (see Boateng et al., 2018; Clark & Watson, 2019) and have instead relied on insufficient procedures to establish or test validity evidence (Juslin & Västfjäll, 2008; Zentner et al., 2008). These issues may be due to a pressing need for practical solutions to conduct research. We argue that this practice is shortsighted and may be hindering the field's ability to compare theories, test hypotheses, and produce meaningful methods to tackle the topic.

Most theories of music and emotion agree that multiple variables – such as music, listener, context, mechanisms, functions, and appraisals – are needed to understand the emotional experiences we undergo while listening to music (Cespedes-Guevara, 2023; Eerola et al., 2025; Juslin, 2025; Scherer & Zentner, 2001). To date, self-report instruments have not adequately captured this dimensionality nor been designed with validity as a primary consideration. Here, we present the start of such a process for a particular theory of music and emotions, which encompasses a broad range of conceptual constructs, including situations, reward, listening modes and agency, social connection, emotion regulation, and musical meanings. We do not claim that the instrument is yet complete, nor that the underlying theory is fully testable in its current form. However, by providing a detailed account of this instrument's development, we aim to set a higher standard for our field pertaining to how psychometric instruments and theories are meant to be carefully aligned.

Acknowledgment

We thank Guoqing Zhu for taking the time to read and evaluate the content.

Action Editor

Andrea Schiavio, University of York.

Peer Review

Melanie Wald-Fuhrmann, Max-Planck-Institut für empirische Ästhetik. Renée Timmers, Sheffield University.

Author Contributions

The authors made the following contributions. Connor Kirts: Conceptualization, Writing–Original Draft, Writing–Review & Editing, Methodology, Data Curation, Investigation, Validation, Visualization, Project Administration; Suvi Saarikallio: Conceptualization, Writing–Original Draft, Writing–Review & Editing, Methodology, Data Curation, Investigation, Validation; Cameron J. Anderson: Data Curation, Investigation, Validation, Writing–Review & Editing; Scott Bannister: Data Curation, Investigation, Validation, Writing–Review & Editing; Julian Céspedes-Guevara, Data Curation, Investigation, Validation, Writing–Review & Editing; Gladys J. Heng: Data Curation, Investigation, Validation, Writing–Review & Editing; Noah Henry: Data Curation, Investigation, Validation, Writing–Review & Editing; Kelly Jakubowski: Data Curation, Investigation, Validation, Writing–Review & Editing; Friederike Koehler: Data Curation, Investigation, Validation, Writing–Review & Editing; Amanda E Krause: Data Curation Investigation, Validation, Writing–Review & Editing; Thomas M. Lennie: Data Curation, Investigation, Validation, Writing–Review & Editing; Isabele Cecilia Martínez, Data Curation, Investigation, Validation, Writing–Review & Editing; Katherine O'Neill, Data Curation, Investigation, Validation, Writing–Review & Editing; Lindsay Warrenburg: Data Curation, Investigation, Validation, Writing–Review & Editing; Tuomas Eerola: Conceptualization, Writing–Original Draft, Writing–Review & Editing, Methodology, Data Curation, Visualization, Investigation, Validation.

Consent for Publication

Not applicable.

Consent to Participate

Participants were informed of study information prior to providing written consent before taking part in the experiment.

Data Availability Statement

All data in support of the research presented here is openly available at https://tuomaseerola.github.io/MIMEE/

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Ethical Approval

Ethical review was obtained by a governing body prior to conducting this research. Approval was granted to conduct this research. Informed consent was obtained prior to participant involvement. This research was granted ethical approval by Durham University (MUS-2024-01-24T18_29_22-bmgm21).

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by the Research Council of Finland (346210) and the European Union (ERC, MUSICONNECT, 101045747).

ORCID iDs

Connor Kirts https://orcid.org/0009-0009-7207-6884
Suvi Saarikallio https://orcid.org/0000-0002-4647-8048
Cameron J. Anderson https://orcid.org/0000-0002-8737-2055
Scott Bannister https://orcid.org/0000-0003-4905-0511
Julian Céspedes-Guevara https://orcid.org/0000-0002-8816-3650
Gladys J. Heng https://orcid.org/0000-0002-0646-2121
Noah Henry https://orcid.org/0000-0002-2384-245X
Kelly Jakubowski https://orcid.org/0000-0002-4954-7117
Friederike Koehler https://orcid.org/0000-0002-0877-7100
Amanda E. Krause https://orcid.org/0000-0003-3049-9220
Thomas M. Lennie https://orcid.org/0000-0001-9821-6173
Isabel Cecilia Martínez https://orcid.org/0000-0002-1837-5957
Katherine O'Neill https://orcid.org/0000-0001-9245-690X
Lindsay Warrenburg https://orcid.org/0000-0002-3986-4573
Tuomas Eerola https://orcid.org/0000-0002-2896-929X

Note

1. "a construct is some postulated property or process characteristic of people, groups, organizations, situations, or environments that is assumed to be reflected in scores on measures of attributes of the property or elements of the process." (Hoyle et al., 2025).

References

Almanasreh, E., Moles, R., & Chen, T. F. (2019). Evaluation of methods used for estimating content validity. *Research in Social and Administrative Pharmacy*, *15*(2), 214–221. https://doi.org/10.1016/j.sapharm.2018.03.066

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. American Educational Research Association.

Baer, R. A., Smith, G. T., & Allen, K. B. (2004). Assessment of mindfulness by self-report: The Kentucky inventory of mindfulness skills. *Assessment*, *11*(3), 191–206. https://doi.org/10.1177/1073191104268029

Baltazar, M., & Saarikallio, S. (2016). Toward a better understanding and conceptualization of affect self-regulation through music: A critical, integrative literature review. *Psychology of Music*, *44*(6), 1500–1521. https://doi.org/10.1177/0305735616663313

Beck, K. (2020). Ensuring content validity of psychological and educational tests – the role of experts. *Frontline Learning Research*, *8*(6), 1–37. https://doi.org/10.14786/flr.v8i6.517

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, *6*, 149. https://doi.org/10.3389/fpubh.2018.00149

Bonneville-Roussy, A., Stillwell, D., Kosinski, M., & Rust, J. (2017). Age trends in musical preferences in adulthood: 1. Conceptualization and empirical investigation. *Musicae Scientiae*, *21*(4), 369–389. https://doi.org/10.1177/1029864917691571

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*(3), 425–440. https://doi.org/10.1007/s11336-006-1447-6

Borsboom, D. (2023). Psychological constructs as organizing principles. In L. A. van der Ark, W. H. M. Emons, & R. R. Meijer (Eds.), *Essays on contemporary psychometrics* (pp. 89–108). Springer International Publishing. https://doi.org/10.1007/978-3-031-10370-4_5

Brown, K. W., & Ryan, R. M. (2003). The benefits of being present: Mindfulness and its role in psychological well-being. *Journal of Personality and Social Psychology*, *84*(4), 822–848. https://doi.org/10.1037/0022-3514.84.4.822

Cespedes-Guevara, J. (2023). A constructionist approach to emotional experiences with music. *Advances in Cognitive Psychology*, *19*(4), 46–62. https://doi.org/10.5709/acp-0402-4

Céspedes-Guevara, J., & Eerola, T. (2018). Music communicates affects, not basic emotions–A constructionist account of attribution of emotional meanings to music. *Frontiers in Psychology*, *9*, 215. https://doi.org/10.3389/fpsyg.2018.00215

Chamorro-Premuzic, T., & Furnham, A. (2007). Personality and music: Can traits explain how people use music in everyday life? *British Journal of Psychology*, *98*(2), 175–185. https://doi.org/10.1348/000712606X111177

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, *31*(12), 1412–1427. https://doi.org/10.1037/pas0000626

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine*, *119*(2), 166.e7–166.e16. https://doi.org/10.1016/j.amjmed.2005.10.036

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/bf02310555

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281–302. https://doi.org/10.1037/h0040957

DeVellis, R. F. (2012). *Scale development: Theory and applications* (3rd ed.). Sage Publications.

Eerola, T (2024). CVI: Content validity indices for psychological research. [Software]. *Github*. https://github.com/tuomaseerola/CVI

Eerola, T., & Kirts, C. (2025). MIMEE. [Software]. *Github*, https://tuomaseerola.github.io/MIMEE/

Eerola, T., Kirts, C., & Saarikallio, S. (2025). Episode model: The functional approach to emotional experiences of music. *Psychology of Music, 53*(4), 590–615. https://doi.org/10.1177/03057356241279763

Epskamp, S., Rhemtulla, M., & Borsboom, D. (2017). Generalized network psychometrics: Combining network and latent Variable models. *Psychometrika*, *82*(4), 904–927. https://doi.org/10.1007/s11336-017-9557-x

Feldman, G., Hayes, A., Kumar, S., Greeson, J., & Laurenceau, J. P. (2007). Mindfulness and emotion regulation: The development and initial validation of the cognitive and affective mindfulness scale-revised (CAMS-R). *Journal of Psychopathology and Behavioral Assessment*, *29*(3), 177–190. https://doi.org/10.1007/s10862-006-9035-8

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid

them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Fowler, F. (1995). *Improving Survey Questions*. Sage Publications Ltd. https://uk.sagepub.com/en-gb/eur/improving-survey-questions/book4994

Groarke, J. M., & Hogan, M. J. (2018). Development and psychometric evaluation of the adaptive functions of music listening scale. *Frontiers in Psychology*, *9*, 516. https://doi.org/10.3389/fpsyg.2018.00516

Gross, J. J., & John, O. P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology*, *85*(2), 348–362. https://doi.org/10.1037/0022-3514.85.2.348

Haucke, M., Hoekstra, R., & van Ravenzwaaij, D. (2021). When numbers fail: Do researchers agree on operationalization of published research? *Royal Society Open Science*, *8*(9), 191354. https://doi.org/10.1098/rsos.191354

Henry, N., Maloney, L., & Egermann, H. (2024). A mixed-method exploratory approach to identifying the utilitarian functions of music listening. *Music & Science*, *7*, https://doi.org/10.1177/20592043241266972

Honnibal, M., Montani, I., Boyd, A., Landeghem, S. V., & Peters, H. (2023). *spaCy: Industrial-strength Natural Language Processing in Python. (v3.7.2)* [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.10009823

Hoyle, R. H., Borsboom, D., & Tay, L. (2025). Measuring constructs. In D. T. Gilbert, S. T. Fiske, E. J. Finkel, & W. B. Mendes (Eds.), *The handbook of social psychology* (6th ed.). Situational Press. https://doi.org/10.70400/OUQF7656

Huron, D. (2006). *Sweet anticipation: Music and the psychology of expectation*. MIT press.

Juslin, P. N. (2013). From everyday emotions to aesthetic emotions: Towards a unified theory of musical emotions. *Physics of Life Reviews*, *10*(3), 235–266. https://doi.org/10.1016/j.plrev.2013.05.008

Juslin, P. N. (2025). Major theories of emotion causation and their applicability to music: The case for multi-level approaches. *Music Perception*, *42*(5), 421–466. https://doi.org/10.1525/mp.2025.2396878

Juslin, P. N., & Laukka, P. (2004). Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, *33*(3), 217–238. https://doi.org/10.1080/0929821042000317813

Juslin, P. N., & Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and Brain Sciences*, *31*(5), 559–575. https://doi.org/10.1017/S0140525X08005293

Kline, P. (2000). *A psychometrics primer* (1st ed.). Free Association Books.

Koehler, F., Silverman, M. J., Riegelman, A., Abbazio, J. M., & Saarikallio, S. (2025). A scoping review and categorization of music and health psychometric inventories. *Psychology of Music*. https://doi.org/10.1177/03057356251322071

Krause, A. E., Mackin, S., Mossman, A., Murray, T., Oliver, N., & Tee, V. (2020). Conceptualizing control in everyday music listening: Defining dominance. *Music & Science*, *3*, https://doi.org/10.1177/2059204320931643

Krause, A. E., North, A. C., & Hewitt, L. Y. (2015). Music-listening in everyday life: Devices and choice. *Psychology of Music*, *43*(2), 155–170. https://doi.org/10.1177/0305735613496860

Kreutz, G., Ott, U., Teichmann, D., Osawa, P., & Vaitl, D. (2008). Using music to induce emotions: Influences of musical preference and absorption. *Psychology of Music*, *36*(1), 101–126. https://doi.org/10.1177/0305735607082623

Law, L. N. C., & Zentner, M. (2012). Assessing musical abilities objectively: Construction and validation of the profile of music perception skills. *PLOS ONE*, *7*(12), e52508. https://doi.org/10.1371/journal.pone.0052508

Lennie, T. M., & Eerola, T. (2022). The CODA model: A review and skeptical extension of the constructionist model of emotional episodes induced by music. *Frontiers in Psychology: Auditory Cognitive Neuroscience*, *13*, 822264. https://doi.org/10.3389/fpsyg.2022.822264

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, *3*(3), 635–694. https://doi.org/10.2466/pr0.1957.3.3.635

Mas-Herrero, E., Marco-Pallares, J., Lorenzo-Seva, U., Zatorre, R. J., & Rodriguez-Fornells, A. (2013). Individual differences in music reward experiences. *Music Perception*, *31*(2), 118–138. https://doi.org/10.1525/mp.2013.31.2.118

Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives*, *15*(2), 51–69. https://doi.org/10.1080/15366367.2017.1348108

Mehr, S. A., Singh, M., Knox, D., Ketter, D. M., Pickens-Jones, D., Atwood, S., Lucas, C., Jacoby, N., Egner, A. A., Hopkins, E. J., Howard, R. M., Hartshorne, J. K., Jennings, M. V., Simson, J., Bainbridge, C. M., Pinker, S., O'Donnell, T. J., Krasnow, M. M., & Glowacki, L. (2019). Universality and diversity in human song. *Science*, *366*, 6468. https://doi.org/10.1126/science.aax0868

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11. https://doi.org/10.3102/0013189X018002005

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika*, *72*(4), 461–473. https://doi.org/10.1007/s11336-007-9039-7

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. https://doi.org/10.1371/journal.pone.0089642

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, *29*(5), 489–497. https://doi.org/10.1002/nur.20147

Randall, W. M., & Rickard, N. S. (2017). Personal music listening: A model of emotional outcomes developed through Mobile experience sampling. *Music Perception: An Interdisciplinary Journal*, *34*(5), 501–514. https://doi.org/10.1525/mp.2017.34.5.501

Rauthmann, J. F. (2015). Structuring situational information. *European Psychologist*, *20*(3), 176–189. https://doi.org/10.1027/1016-9040/a000225

Revelle, W. (2024). The seductive beauty of latent variable models: Or why I don't believe in the easter bunny. *Personality and Individual Differences*, *221*, 112552. https://doi.org/10.1016/j.paid.2024.112552

Romero Jeldres, M., Díaz Costa, E., & Faouzi Nadim, T. (2023). A review of Lawshe's method for calculating content validity in the social sciences. *Frontiers in Education*, *8*, 1271335. https://doi.org/10.3389/feduc.2023.1271335

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, *76*(5), 805–819. https://doi.org/10.1037/0022-3514.76.5.805

Saarikallio, S. H. (2008). Music in mood regulation: Initial scale development. *Musicae Scientiae*, *12*(2), 291–309. https://doi.org/10.1177/102986490801200206

Saarikallio, S. H. (2012). Development and validation of the brief music in mood regulation scale (B-MMR). *Music Perception: An Interdisciplinary Journal*, *30*(1), 97–105. https://doi.org/10.1525/mp.2012.30.1.97

Saarikallio, S. H., Randall, W. M., & Baltazar, M. (2020). Music Listening for Supporting Adolescents' Sense of Agency in Daily Life. *Frontiers in Psychology*, *10*, 20911. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.02911 https://doi.org/10.3389/fpsyg.2019.02911

Sandstrom, G. M., & Russo, F. A. (2013). Absorption in music: Development of a scale to identify individuals with strong emotional responses to music. *Psychology of Music*, *41*(2), 216–228. https://doi.org/10.1177/0305735611422508

Schäfer, T., Sedlmeier, P., Städtler, C., & Huron, D. (2013). The psychological functions of music listening. *Frontiers in Psychology*, *4*, 511. https://doi.org/10.3389/fpsyg.2013.00511

Scherer, K. R., & Coutinho, E. (2013). How music creates emotion: A multifactorial process approach. In *The emotional power of music: Multidisciplinary perspectives on musical arousal, expression, and social control* (pp. 121–145). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199654888.003.0010

Scherer, K. R., & Zentner, M. R. (2001). Emotional effects of music: Production rules. In P. N. Juslin & J. A. Sloboda (Eds.), *Music and emotion: Theory and research* (pp. 361–392). Oxford University Press.

Schindler, I., Hosoya, G., Menninghaus, W., Beermann, U., Wagner, V., Eid, M., & Scherer, K. R. (2017). Measuring aesthetic emotions: A review of the literature and a new assessment tool. *PloS ONE*, *12*(6), e0178899. https://doi.org/10.1371/journal.pone.0178899

Sireci, S. G. (1998). The construct of content validity. *Social Indicators Research*, *45*(1), 83–117. https://doi.org/10.1023/A:1006985528729

Thompson, W. F., Bullot, N. J., & Margulis, E. H. (2023). The psychological basis of music appreciation: Structure, self, source. *Psychological Review*, *130*(1), 260–284. https://doi.org/10.1037/rev0000364

van Ede, F., Board, A. G., & Nobre, A. C. (2020). Goal-directed and stimulus-driven selection of internal representations. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(39), 24590–24598. https://doi.org/10.1073/pnas.2013432117

Vigl, J., Ojell-Järventausta, M., Sipola, H., & Saarikallio, S. (2023). Melody for the mind: Enhancing mood, motivation, concentration, and learning through music listening in the classroom. *Music & Science*, *6*, https://doi.org/10.1177/20592043231214085

Waltz, C. F., Strickland, O. L., & Lenz, E. R. (2005). *Measurement in nursing and health research* (3rd ed.). Springer Publishing Co.

Wanzer, D. L., Finley, K. P., Zarian, S., & Cortez, N. (2020). Experiencing flow while viewing art: Development of the aesthetic experience questionnaire. *Psychology of Aesthetics, Creativity, and the Arts*, *14*(1), 113–124. https://doi.org/10.1037/aca0000203

Warrenburg, L. A. (2020). Comparing musical and psychological emotion theories. *Psychomusicology: Music, Mind and Brain*, *30*(1), 1–19. https://doi.org/10.1037/pmu0000247

Weining, C. (2022). Listening modes in concerts: A review and conceptual model. *Music Perception*, *40*(2), 112–134. https://doi.org/10.1525/mp.2022.40.2.112

Zentner, M., Grandjean, D., & Scherer, K. R. (2008). Emotions evoked by the sound of music: Characterization, classification, and measurement. *Emotion*, *8*(4), 494–521. https://doi.org/10.1037/1528-3542.8.4.494