frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 28 November 2022
DOI 10.3389/feduc.2022.1013841

OPEN ACCESS

EDITED BY
Gavin T. L. Brown,
The University of Auckland, New Zealand

REVIEWED BY
Quan Zhang,
Jiaxing University,
China
Anna Wolf,
University of Hamburg,
Germany

*CORRESPONDENCE
Julia Ehninger
info@juliaehninger.de
Jens Knigge
jens.knigge@nord.no

†These authors share first authorship

SPECIALTY SECTION
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

RECEIVED 07 August 2022
ACCEPTED 28 October 2022
PUBLISHED 28 November 2022

CITATION
Ehninger J, Knigge J and Rolle C (2022) Why are certain items more difficult than others in a competency test for music–related argumentation?
Front. Educ. 7:1013841.
doi: 10.3389/feduc.2022.1013841

Why are certain items more difficult than others in a competency test for music–related argumentation?

Julia Ehninger[1]*†, Jens Knigge[2]*† and Christian Rolle[1]

[1]Institute for Music Education, University of Cologne, Cologne, Germany, [2]Department for Arts and Culture, Nord University, Levanger, Norway

This paper examines why certain items in a competency test for music-related argumentation are more difficult than others. Based on previous studies on school-related achievement tests, the authors assume that differences in item difficulty are related to different item characteristics or combinations of characteristics. In this study, the item characteristics of a test for music-related argumentation were first identified and coded. Three domains were identified as contributing to item difficulty: cognitive requirements, knowledge, and formal item features. Second, multiple linear regression analyses were conducted with the item characteristics as predictors of item difficulty, which had been estimated in a prior study. A comparison of three regression models confirmed that the model holding four predictors of the domain "cognitive requirements" best fit the study data $R_{adj}{}^2 = 0.71.$ The strongest predictor in the final model was "reference to musical attributes" $(\beta = 0.46, \beta = 0.51),$ followed by "cross-sentence argumentation" $(\beta = 0.37)$ and "dialogical argumentation" $(\beta = 0.20).$ These results indicate that the difficulty of an item increased most when participants had to refer to musical attributes to solve the task. The items that required the participants to provide cross-sentence or dialogical argumentation were more challenging as well. The findings regarding the relations between item characteristics and item difficulty contribute to a better understanding of music-related argumentative competence, with important implications for the music classroom.

KEYWORDS

assessment and education, competency testing, music education, music–related argumentation, item characteristics, empirical research

Introduction

Argumentation is an essential part of our everyday lives. We are familiar with it from discussions at work and debates in court, and it is an integral part of the democratic process. Argumentation also plays a major role—whether intended or not—in the classroom. Students require argumentative competence in order to engage in classroom activities such as group discussions. Furthermore, discourse practices such as arguing and explaining contribute to the acquisition and negotiation of knowledge (Kuhn, 2005; Morek and Heller, 2012; Morek et al., 2017). Therefore, it is not surprising that argumentative skills are considered a key competency for students' overall educational success (Quasthoff et al., 2020b). Although it has become evident in recent years that language is constitutive of learning in all school subjects (Lazarou et al., 2016; Rapanta, 2018; Quasthoff et al., 2020a), there has been little theoretical and empirical research on the role of language competence, such as argumentative competence, in music as a school subject (Bossen, 2017).

Language is an important medium of communication, including in the music classroom. When rehearsing, musicians often verbally negotiate how music should sound, whether it is a band working on a song or members of a string quartet who must agree on the interpretation of the musical piece they are rehearsing. In music lessons, verbal engagement with music plays a major role and music-related argumentative competence is an integral part of German school curricula (Kultusministerkonferenz, 2005). In a prior study, we empirically modeled music-related argumentative competence and developed the MARKO test, a competency test for *music-related argumentative competence* (Musikbezogene ARgumentationsKOmpetenz; German for music-related argumentative competence).

In this paper, we explore the question of why certain items of the competency test were more difficult than others. Prior research on competency tests has shown that certain item characteristics can increase an item's difficulty (e.g., Knigge, 2010). For example, items containing a great deal of text can be more challenging for students to solve (e.g., Prenzel et al., 2002). The complexity of a musical piece can also contribute to the difficulty of an item (e.g., Knigge, 2010, pp. 228–231). Therefore, we present in-depth analyses on the item characteristics of the items of the MARKO test for music-related argumentative competence. We analyze which item characteristics contribute to item difficulty and closely examine the requisite competencies for solving the items in the competency test.

Theoretical background

Music–related argumentative competence

Music-related argumentative competence can be defined as the "context-specific cognitive disposition that is acquired and needed to justify and defend esthetic judgments about music in a comprehensive, plausible, and differentiated way" (Ehninger et al., 2021, pp. 2–3). The competence to reflect on and justify judgments about music is relevant in school curricula in many countries (e.g., Germany: Kultusministerkonferenz, 2005; Norway: Utdanningsdirektoratet, 2020); however, until now, there has been little research on the requirements for engaging in music-related argumentation.

Rolle (2013) proposed a theoretical competency model on music-related argumentation and distinguished between several competency levels. This model assumes that it is easier to refer to subjective impressions of music and personal taste than the cultural and social context of music or esthetic conventions. People on higher competency levels are better able to reflect on their own judgment about music and can integrate criticism and other people's opinions into their reasoning (see also Knörzer et al., 2016). Based on Rolle's theoretical assumptions, the MARKO competency test for music-related argumentation was developed and validated.

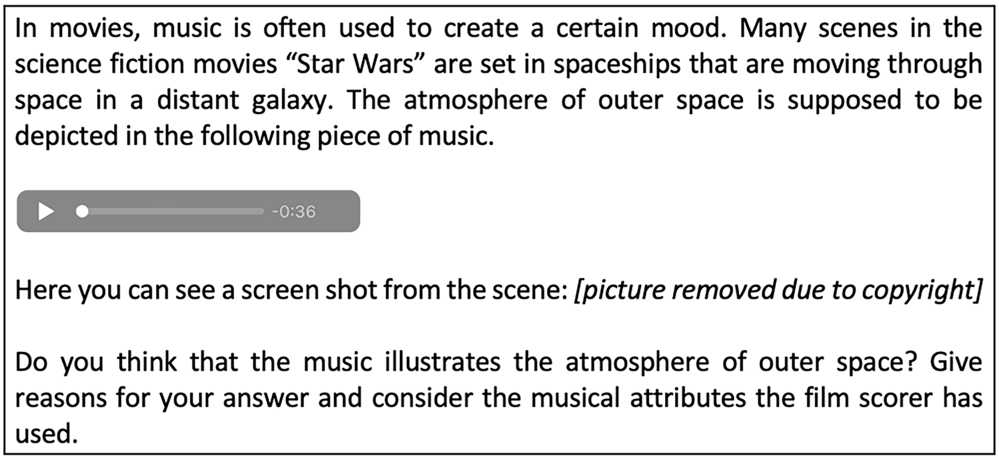
Competency test for music-related argumentation (MARKO)

The MARKO test is in German and includes 25 open-ended items distributed online. It was designed for ninth to twelfth grade high school students as well as university students. During the test, the participants worked individually on computers and used headphones to listen to music (and sometimes watch videos) of various musical genres. In the test, they were asked to justify their esthetic judgment in a written answer. For example, they were asked why they thought a musical piece created a certain atmosphere or were prompted to comment on a discussion below a YouTube video or a concert review in a newspaper.

The validation of the test as well as a competency model resulting from data collected from 440 participants were presented in a prior study (Ehninger et al., 2021; Ehninger, 2022). Two sample items of the test are presented below to provide an insight into the test. Figure 1 shows the sample item "Star Wars," which was developed to assess how the participants referred to the atmosphere and musical attributes of a musical piece. In this item, the participants were asked whether a musical piece illustrated the atmosphere in outer space. To solve the item, the participants produced texts that were rated in accordance with a coding scheme (Table 1).

While some items in the test were aimed at assessing how the participants referred to musical attributes or their subjective impressions of the music, others were designed to measure the dialogical dimension of argumentation. In the item "Eurovision Song Contest," the participants were asked to comment on a discussion on YouTube about the winner of the *Eurovision Song Contest* (Figure 2). The participants' answers were also coded with a coding scheme (Table 2).

The two sample items differed in various respects. While the "Star Wars" item (Figure 1) contained little text, the participants had to read a great deal of text before solving the "Eurovision Song Contest" item (Figure 2). Furthermore, the cognitive requirements for solving the items seemed to differ. For the "Eurovision Song Contest" item, the participants had to consider the social and cultural contexts, such as feminism and social justice. In comparison to the "Eurovision Song Contest" item, a much more differentiated reference to musical attributes was required in the



In movies, music is often used to create a certain mood. Many scenes in the science fiction movies "Star Wars" are set in spaceships that are moving through space in a distant galaxy. The atmosphere of outer space is supposed to be depicted in the following piece of music.

Here you can see a screen shot from the scene: *[picture removed due to copyright]*

Do you think that the music illustrates the atmosphere of outer space? Give reasons for your answer and consider the musical attributes the film scorer has used.

FIGURE 1
Test item "Star Wars" (English translation). Note: The participants listened to an excerpt from the film score (Arrival at Naboo, Episode I). A screenshot from the scene was shown in the item but had to be omitted here due to copyright concerns. The screenshot showed the view of a planet from a space shuttle cockpit (see also Ehninger et al., 2021; all items are available here: DOI 10.17605/OSF.IO/ZVP4B).

"Star Wars" item, at least for scoring the maximum number of points.

These two example items show that the requirements for solving an item (category) could vary in a competency test and that the content of the items could also differ. However, we did not know why one item (category) was more difficult than the other. Was the "Eurovision Song Contest" item more difficult because it contained more text, or was the "Eurovision Song Contest" item perhaps easier because the students both listened to music and watched a video? These questions can be answered by examining the characteristics of the items and relating them to the difficulty of the items.

Item difficulty and item characteristics

Item characteristics can be defined as the characteristics of an item associated with higher or lower demands on test takers, thereby influencing the solution probability (Hartig and Jude, 2007, p. 31). They are relevant for competence research primarily because the "competence" construct is defined by its context-specificity (whereas, e.g., intelligence is defined as generalized, context-independent cognitive dispositions that can only be learned to a limited extent; see, e.g., Hartig and Klieme, 2006; Hartig, 2008). From this context-specificity, one can derive the fundamental interest in the characteristics of a situation (i.e., the item in a test situation) in which competent performance manifests itself. Particular attention is paid to the characteristics of a situation that make competent performance easier or more difficult. This is because only "knowledge of the situational characteristics that influence successful performance enables a deeper understanding of the processes that underlie successful performance and thus a better understanding of the competence in question" (Hartig and Jude, 2007, p. 31; translation by the authors).

Nevertheless, there are also other arguments regarding the relevance of item characteristics: One interesting aspect is that they can be used to define levels of competence (e.g., Hartig, 2007). If different item difficulties can be explained empirically by a certain set of item characteristics, the levels of a competency can be described by means of the characteristics in question. These competency-level descriptions are then empirically validated and are also generalizable beyond the concrete test items used (Hartig and Jude, 2007).

Another aspect concerns the validity of a test. In their influential paper, Borsboom et al. (2004) argue for a reconceptualization of test validity: "A test is valid for measuring an attribute if (a) the attribute exists and (b) variations in the attribute causally produce variation in the measurement outcomes" (p. 1061). Therefore, validation research must be directed "at the processes that convey the effect of the measured attribute on the test scores" (p. 1061). Against this background, the formulation of item characteristics can be understood as hypotheses about the processes that cause variation in a competency test. Hence, from a test-theoretical point of view, the prediction of item difficulty by item characteristics can be regarded as a confirmation of the validity of the measurement instrument (see also Hartig, 2007).

Furthermore, if empirically validated item characteristics are provided, they can be used to design new test items (Nold and Rossa, 2007). It would then be possible to create specific "requirement profiles" for the items that are supposed to be developed, which would consist of different combinations and degrees of the item characteristics. Model-guided item development, in this sense, makes it possible to determine *a priori* which items should be easier or more difficult and the reasons for

TABLE 1 Coding scheme for the sample item "star wars."

Points	Description	Sample answers
0	Tautological justification or no reason	"Yes, because of the atmosphere that exists in space. The composer presented this very well." (VP_661)
1	Participants refer only to the musical atmosphere. If musical attributes are mentioned (or even a causal relationship is established between them and the atmosphere), this is done by referring to "basic" and superficial characteristics of the music (e.g., "bright notes," "long tones," "loud," "soft," "instruments that create tension").	"I think so because it sounds exciting and unusual, which, in my opinion, corresponds well with the atmosphere in outer space." (VP_714)
2	Participants relate the generated atmosphere to musical attributes. If instruments (e.g., "quiet strings") are mentioned, the answer is given two points.	"Yes, I find it very well done. The sound layers depict the infinite vastness of the universe … the synthesizers give the piece a futuristic character … single high notes to illustrate the stars." (VP_589)
3	Participants relate the generated atmosphere to musical attributes. A detailed description is provided (e.g., the musical form and the way the instruments are played).	"I find the composition convincing because the long notes (played by the violin) generate a feeling of width and yet (because of the high notes) sound quite exciting and dramatic, especially at the beginning. The fast (xylophone?) notes that go up and down the scale have a bright sound and are reminiscent of stars. The flourish at the beginning could suggest that a scenery of spectacular surroundings is just revealing itself to the audience." (VP_610)

This is a simplified and condensed version of the coding scheme, which was also published in Ehninger et al. (2021).

these differences. Accordingly, items can be developed explicitly for a certain competence profile or competence level.

Prenzel et al. (2002, p. 125) proposed categorizing item characteristics into three domains: formal task characteristics, cognitive demands in solving the tasks, and the characteristics of the knowledge base required for solving the tasks (similar categorizations can be found in, e.g., Nold and Rossa, 2007, and Hartig and Klieme, 2006). Knigge (2010) used this systemization for a music-specific item analysis of a competency test for musical perception (KoMus test). He systematized the item characteristics as follows: (1) formal item characteristics, (2) cognitive demands on auditory perception and musical memory, and (3) necessary activation of expertise.

(1) *Formal item* characteristics include the item format (closed vs. open), the formalities of the item content (e.g., picture stimulus vs. auditory stimulus), and the nature of the item stem (e.g., long vs. short question phrases). The influence of this item characteristic domain has been demonstrated in studies on the assessment of language and mathematical/scientific competencies (e.g., Prenzel et al., 2002; Cohors-Fresenborg et al., 2004; Beck and Klieme, 2007). With regard to musical competence, an influence of such general, non-music-specific characteristics also seems plausible, which was also confirmed by Knigge (2010) and Jordan (2014).

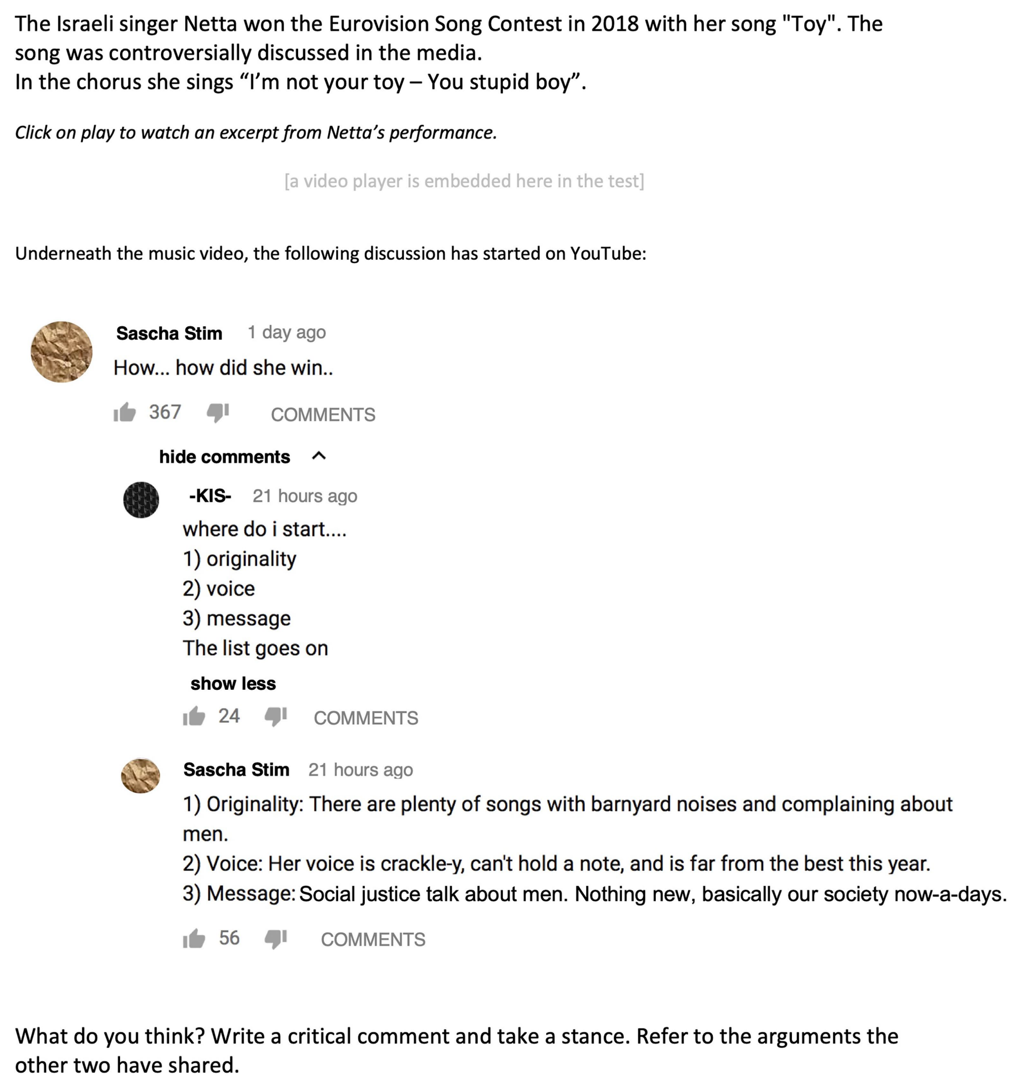
(2) There are several research results from other disciplines regarding the requisite cognitive processes for processing an item in language or mathematical tests (e.g., Hartig and Klieme, 2006; Nold and Rossa, 2007); however, these results are not directly transferable to musical competence. With regard to a competence test for musical perception, Knigge (2010) identified two cognitive demand domains of relevance to item processing related to auditory perception and musical memory. While these requirements are closely connected, they can also occur independently of each other.

(3) Finally, item characteristics can be characterized by the activation of subject-specific knowledge (e.g., Prenzel et al., 2002). In relation to the musical knowledge required to solve an auditory perception item, Knigge (2010) identified five item characteristics: knowledge of musical notation, knowledge of music theory, knowledge of music history, knowledge of musical styles and genres, and knowledge of the cultural and social contexts of music.

The categorization of item characteristics presented above was empirically validated by Knigge (2010) and Jordan (2014) who showed that the difficulty of an item was mainly influenced by the cognitive demands on auditory perception and necessary subject knowledge. For the entire set of item characteristics, a very strong prediction of item difficulty could be demonstrated for the KoMus competency test, with the explained variance being between 55% and 83% (regression analyses were conducted for all four subdimensions of the KoMus test; Jordan, 2014, pp. 136–139).

Research goal

The aim of this paper was to explore the question of why certain items of the MARKO test for music-related argumentation were more difficult than others. Based on previous studies on school-related achievement tests, we assumed that the differences in item difficulty were related to different item characteristics or combinations of



The Israeli singer Netta won the Eurovision Song Contest in 2018 with her song "Toy". The song was controversially discussed in the media.
In the chorus she sings "I'm not your toy – You stupid boy".

Click on play to watch an excerpt from Netta's performance.

[a video player is embedded here in the test]

Underneath the music video, the following discussion has started on YouTube:

Sascha Stim 1 day ago
How... how did she win..
👍 367 👎 COMMENTS

hide comments ^

-KIS- 21 hours ago
where do i start....
1) originality
2) voice
3) message
The list goes on
show less
👍 24 👎 COMMENTS

Sascha Stim 21 hours ago
1) Originality: There are plenty of songs with barnyard noises and complaining about men.
2) Voice: Her voice is crackle-y, can't hold a note, and is far from the best this year.
3) Message: Social justice talk about men. Nothing new, basically our society now-a-days.
👍 56 👎 COMMENTS

What do you think? Write a critical comment and take a stance. Refer to the arguments the other two have shared.

FIGURE 2
Sample item "Eurovision Song Contest" (English translation). Note: A short excerpt from Netta's performance was embedded (1,03–1,40). The video of her whole performance can be found at: https://www.youtube.com/watch?v=84LBjXaeKk4 (see also Ehninger et al., 2021).

characteristics. Therefore, our aim was to identify the item characteristics relevant to the MARKO test and quantify their specific influence. In doing so, we hoped to gain a better understanding of the specific competence needed to solve the competency test items as well as examine the validity of the test.

Materials and methods

The methodological approach in this paper can be divided into three steps: First, the item characteristics were identified and categorized ("Identification and Categorization of Item Characteristics"). Second, the whole item pool (i.e., competency test) was coded according to the identified and categorized item characteristics ("Coding Item Characteristics"). Finally, multiple linear regression analyses were conducted with the item characteristics as predictors of item difficulty ("Multiple Regression Analyses"). The difficulty parameters of the test items were obtained in a prior study ($N = 440$; students from upper secondary schools and universities) employing IRT scaling (partial credit model; Ehninger et al., 2021).

Identification and categorization of item characteristics

We chose a combined deductive–inductive approach to identify item characteristics specific to the MARKO test:

- We adapted findings from previous research on item characteristics from music and other subjects (e.g., Knigge, 2010; Jordan, 2014).

TABLE 2 Coding scheme for the item "Eurovision song contest" (English translation).

Points	Coding scheme	Sample answers
0	The answer paraphrases parts of the YouTube discussion and/or refers to personal taste.	"I do not like the song either and agree with Sascha's comment. I also think that she does not hold the pitch very well, and I think that the crackling is a little ridiculous." (P2_11)
1	The answer takes into account the entire YouTube discussion but is paraphrasing it for the most part. The answer might include a new argument that has not come up in the YouTube discussion.	"The singer addresses a very important and current topic: social equality. However, I find it is not really appropriately communicated. The lyrics are presented with humor and thus they do not mean anything." (VP_142)
2	The answer contains at least two new arguments. Different perspectives are evaluated.	"Women's empowerment is a current topic of great importance. It is good that artists are setting an example. Sometimes, the lyrics are one-dimensional because women also 'play' with women. But often, it is the other way around and has been the case for centuries due to the unfair distribution of power, where women are neglected. Maybe she should have sung 'I'm not a toy, for no one' or something like that, which emphasizes the idea of equality. She represents a strong image of women, which is definitely socially critical. Because of the 'crackling,' as Sascha calls it, the song is unusual and different and differs from the social norm that influences the masses, as Sascha and 367 other people show. Have fun with your followers and mainstream boredom." (VP_89)

- We used the MARKO coding schemes and the MARKO competency model (Ehninger et al., 2021).
- We took theoretical assumptions on musical perception (see "Domain 1: cognitive requirements") and argumentative competence into account (e.g., Heller and Morek, 2015; see "Domain 1: cognitive requirements").
- We conducted in-depth analyses of the individual items.

Against this background, we conducted several coding sessions in which the applicability of the identified item characteristics was tested for the entire item pool (the MARKO test consists of 25 polytomous items). The coding sessions were conducted in a circular procedure that was carried out several times until interrater-reliability was acceptable. In those sessions, item characteristics were first coded by two independent raters for every single item category of the test. In a second step, interrater-reliability was calculated and ratings with low interrater-reliability were reviewed. Next, item characteristics were revised, and new item characteristics were added if necessary. Finally, all item categories were rated again. On this basis, several item characteristics were identified for the item pool of the MARKO test, resulting in three domains of item characteristics: *(1) cognitive requirements, (2) knowledge,* and *(3) formal item features.*

Domain 1: Cognitive requirements

The first domain of the identified item characteristics dealt with the cognitive requirements that a participant had to cope with when solving an item. Many cognitive requirements were described in the coding schemes for every item (Tables 1, 2). Three characteristics were identified as dealing with cognitive requirements: *(a) reference to perceived musical attributes, (b) cross-sentence argumentation,* and *(c) dialogical argumentation* (Table 3) .

A *reference to perceived musical attributes* was required for many items and was specified in the coding schemes. Research on music-related argumentation has shown that references to the musical attributes of a musical piece are a cognitive operation that is essential when engaging in music-related argumentation (Rolle, 2013; Knörzer et al., 2016). For example, the coding scheme of the "Star Wars" item (Table 1) specified that the participants only had to refer to "basic and superficial characteristics of the music (e.g., 'bright notes', 'long tones' [...])" in order to score one point. To achieve two or more points for the item, the participants had to refer to more specific musical attributes such as musical instruments or the musical form. This item characteristic was also identified in an assessment test on music-related perception (Knigge, 2010; Jordan et al., 2012) and can be framed inside cognitive research in music psychology. From a cognitive psychology perspective, musical perception can be described as the active (re)construction of auditory events with the help of specific techniques and using existing knowledge that is strongly culturally influenced (Morrison and Demorest, 2009; for an overview of findings on musical perception in cognitive neuroscience see Koelsch, 2019). In general, we assume that if a MARKO item demands more complex musical perception, this will lead to an increase in item difficulty.

The item characteristic *cross-sentence argumentation* points to linguistic requirements for producing an answer to an item. The needed discourse competence has been modeled as a dimension of the overarching communicative competence (Canale and Swain, 1980). When people engage in argumentation, they do not "communicate with each other by simply producing words and sentences but by orienting to and accomplishing discursive activities above the sentence-level" (Heller and Morek, 2015, p. 181). In considering the item "Eurovision Song Contest" and its coding scheme (Figure 2; Table 2), it became clear that to score two points for the item, the reasoning of the participant had to be consistent across several sentences.

In the MARKO test, several items were designed to assess the *dialogical* dimension of *argumentation* acknowledging that argumentation must not only be seen as a relationship between sentences but is a social practice (Eemeren et al., 2014, chapter 10). For this reason, in several items, the participants were confronted with opinions of others. An example of cognitive requirement was evident in the "Eurovision Song Contest" item (Figure 2) where the participants had to consider another perspective—an item characteristic called *dialogical argumentation*.

Domain 2: Knowledge

The second domain of item characteristics was entitled "knowledge," which included the item characteristics *(d) cultural and social context of music* and *(e) familiarity of musical genre* (Table 3). Similar item characteristics (knowledge of music history, knowledge of musical styles and genres, and knowledge of cultural and social contexts of music) were investigated and empirically validated by Knigge (2010) and Jordan (2014).

In accordance with Rolle's (2013) theoretical competency model, some items of the MARKO test included information about the *cultural and social contexts of music*. The YouTube discussion around the item "Eurovision Song Contest" (Table 2) referenced "women's empowerment" and "social justice." To understand these references, the participants had to know about the respective discourses and be familiar with them.

The second item characteristic in this domain was *familiarity of musical genre.* The test items included music from various musical genres, such as classical music, pop, musical theater, and hip-hop. The degree to which the participants were familiar with different musical genres varied considerably. For several musical pieces presented in the test, the students provided information on their familiarity with a specific kind of music. This item characteristic captured whether the participants were familiar with the type of music presented in the item. Here, we hypothesized that a person who is familiar with a music genre has more knowledge about this genre and is, therefore, more likely to be able to solve a respective item. Therefore, this item characteristic should lessen the difficulty.

Domain 3: Formal item features

The third domain of item characteristics involved *formal item features* and included item characteristics dealing with the content of the item: *(f) text length, (g) linguistic demands,* and *(h) visuals*.

For the item characteristic *text length*, it became clear that a comparison of the two sample items illustrated earlier (Figures 1, 2) led to significant differences in the amount of text that the participants had to read in order to solve the item. Text length was also identified by Knigge (2010, p. 209) as a difficulty-increasing item characteristic.

The item characteristic *linguistic demands* referenced the vocabulary and grammatical structure used in an item. Nold and Rossa (2007) and Knigge (2010, p. 209) also identified linguistic demands as a difficulty-increasing item characteristic. While all items included the music that the participants were listening to, some items also contained a video or picture. This formal item feature was represented by the item characteristic *visuals*.

Coding item characteristics

Following the identification of the item characteristics, all polytomous item categories were coded. This coding process is exemplified in Figure 2 through the "Eurovision Song Contest" item. This item had two item categories because the participants' answers were rated with 0, 1, or 2 points (Table 2). If a person received one point for the item, they solved item category one, and if they received two points, they solved item category two. While the item characteristics of *Domain 2 (knowledge and familiarity;* Table 4) and the *formal item features (Domain 3*; Table 5) were the same for item categories one and two, *the cognitive requirements (Domain 1;* Table 3) differed between the item categories.

Table 6 shows the item characteristics for both item categories. While there was no need to refer to complex *musical attributes (a)*, consistency in reasoning *across several sentences (b)* was required

TABLE 3 Domain 1 of the identified item characteristics (predictors classified as the cognitive requirements for solving the item).

Domain 1: Cognitive requirements

Predictor	Code	Description
(a) Reference to musical attributes	0	To solve the item, only a reference to salient musical attributes (e.g., "loud," "soft," "long tones") or no musical attribute is necessary.
	1	To solve the item, a reference has to be made to musical attributes that are more complex than salient musical attributes.
	2	To solve the item, several musical attributes have to be named precisely.
(b) Cross-sentence argumentation	0	No elaborate reasoning is needed to solve the item.
	1	Reasoning has to be consistent across several sentences.
(c) Dialogical argumentation	0	There is no need to discuss different perspectives or opinions in the answer.
	1	To solve the item, participants have to take into account different opinions and perspectives on the presented musical piece.

TABLE 4 Domain 2 of the identified item characteristics (predictors classified as "knowledge").

Domain 2: Knowledge		
Predictor	**Code**	**Description**
(d) Cultural and social context of music	0	The social and cultural context of music is not relevant to the item.
	1	The social and cultural context of music is a central element of the item.
(e) Familiarity of musical genre	0	Participants are unfamiliar or somewhat familiar with the type of music heard in the task.
	1	Participants are very familiar with the type of music heard in the task.

TABLE 5 Domain 3 of the identified item characteristics (predictors classified as formal task features).

Domain 3: Formal item features		
Predictor	**Code**	**Description**
(f) Text length	0	Item contains little text
	1	Item contains a lot of text
(g) Linguistic demands	0	Vocabulary: use of high-frequency words Grammar: simple syntactic structures (parataxis, avoidance of complex structures)
	1	Vocabulary: less frequent words, extended vocabulary Grammar: more complex structures
(h) Visuals	0	Item does not include visuals (video/picture).
	1	Item does include visuals (video/picture).

TABLE 6 Coded item characteristics for the item "Eurovision song contest."

Domain	**1 Cognitive requirement**			**2 Knowledge**		**3 Formal item features**		
Item characteristic	***a***	***b***	***c***	***d***	***e***	***f***	***g***	***h***
Item category 1	*0*	*1*	*0*	*1*	*1*	*1*	*1*	*1*
Item category 2	*0*	*1*	*1*	*1*	*1*	*1*	*1*	*1*

for both item categories. *Dialogical argumentation (c)* was not needed to score one point for the "Eurovision Song Contest" item. However, it was required for receiving two points since other people's opinions had to be referenced.

Multiple regression analyses

Multiple regression analyses were conducted in the final step. Here, the item characteristics were used to predict item difficulty. The item difficulty parameters had been estimated with IRT scaling in a prior study (weighted likelihood estimation; Ehninger et al., 2021), where the collected test data were modeled as a partial credit model, and threshold parameters τ were estimated. This presented item difficulty for each item category, with a higher τ value indicating a more difficult item category.

In the multiple regression analyses, the dummy coded item characteristics were used to predict item difficulty τ. In the equation below, τ_i stands for the item difficulty parameter τ of item i. β_0 represents the regression constant and β_c the regression weight for the item characteristic c. Finally, x_{1i} is the code for an item characteristic (1 if the characteristic was present in the item, 0 if it was not).

$$\tau_i = \beta_0 + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \cdots + \beta_c \cdot x_{ci}$$

The difficulty τ of each item category was modeled as the weighted sum of the item characteristics present in a given item category. The regression weights β_c represented the magnitude of influence of an item characteristic on item difficulty. Thus, an item with the characteristic c was β_c more difficult than an item without this item characteristic.

It was assumed that Domain 1 (cognitive requirements) would have a greater impact on item difficulty than the characteristics of the two other domains. Thus, three regression models were estimated. The first model was estimated with predictors from Domain 1 (*cognitive requirements*), the second with predictors from Domain 1 and Domain 2 (*cognitive requirements* and *knowledge*), and the third with the predictors from all three domains. The three models were then compared to one another, and the analyses were conducted in *R* (version 4.1.2). We also checked several assumptions of our data, such as homoscedasticity and multicollinearity. Both the beta coefficients and the collinearity statistics had to be acceptable (VIF < 10; variance inflation factor). In addition, we analyzed the standard errors of the regression coefficients and the part and partial correlations of each predictor variable.

Results

All the item characteristics were dummy coded in preparation for the multiple linear regression analyses. The two-factor variable "reference to musical attributes" had to be converted into two dummy variables ("reference to musical attributes 1" and "reference to musical attributes 2"). All the item characteristics were rated by two raters, who agreed to a great extent ($\kappa\left[0.71,1\right]$).

Next, block-wise multiple linear regression analyses were conducted. The first model included predictors from Domain 1 (*cognitive requirements*); the second model included item characteristics from Domain 1 and Domain 2 (*cognitive requirements* and *knowledge*); and the third model yielded all item characteristics from all three domains (*cognitive requirements*, *knowledge*, and *formal task features*).

Table 7 shows the results of the regression analyses. All four predictors in Model 1 were significant. The reference to musical attributes was the strongest predictor $\left(\beta = 0.46, \beta = 0.51\right)$, followed by cross-sentence argumentation $\left(\beta = 0.37\right)$ and dialogical argumentation $\left(\beta = 0.20\right)$, ${R_{adj}}^2 = 0.71$. The predictors added in Model 2 and Model 3, however, lay above the significance level $\left(p > 0.05\right)$ and roughly explained the same observed variance in item difficulty as in Model 1 $\left({R_{adj}}^2 = 0.71\right)$. A model comparison confirmed that Model 2 did not explain more variance than Model 1, $F\left(2,46\right) = 1.24, p = 0.30$, and that Model 3 did not explain more variance than Model 2, $F\left(3,43\right) = 0.73, p = 0.54$. Therefore, Model 1 suited our data best and met the assumption of non-multicollinearity ($VIF\left[1.08,2.04\right]$). Figure 3 shows four residual plots illustrating the model specification, the normal distribution of the residuals, the

TABLE 7 Regression results with the criterion of item difficulty (thresholds).

Predictor		*b*	*SE*	*beta*	*beta* 95% CI [LL, UL]	*p*	Fit
Model 1: Cognitive requirements							
(Intercept)		−0.60	0.19			<0.01	
(a)	Musical attributes 1	1.80	0.31	0.46	[0.30, 0.61]	<0.001	
	Musical attributes 2	2.55	0.51	0.51	[0.31, 0.72]	<0.001	
(b)	Cross-sentence argumentation	1.45	0.42	0.37	[0.15, 0.58]	<0.01	
(c)	Dialogical argumentation	1.54	0.69	0.20	[0.02, 0.38]	0.03	
							${R_{adj}}^2 = 0.71^{**}$
Model 2: Cognitive requirements and knowledge							
(Intercept)		−0.48	0.20			0.02	
(a)	Musical attributes 1	1.83	0.31	0.46	[0.30, 0.62]	<0.001	
	Musical attributes 2	2.62	0.51	0.53	[0.32, 0.73]	<0.001	
(b)	Cross-sentence argumentation	1.42	0.43	0.36	[0.14, 0.58]	<0.01	
(c)	Dialogical argumentation	1.59	0.78	0.21	[0.00, 0.41]	<0.05	
(d)	Context of music	0.18	0.49	0.04	[−0.16, 0.24]	0.72	
(e)	Familiarity	−0.49	0.31	−0.13	[−0.29, 0.04]	0.12	
							${R_{adj}}^2 = 0.71^{**}$
Model 3: Cognitive requirements and knowledge and formal item features							
(Intercept)		−0.73	0.27			<0.01	
(a)	Musical attributes 1	1.90	0.32	0.48	[0.32, 0.64]	<0.001	
	Musical attributes 2	2.73	0.52	0.55	[0.34, 0.76]	<0.001	
(b)	Cross-sentence argumentation	1.43	0.43	0.36	[0.14, 0.58]	<0.01	
(c)	Dialogical argumentation	1.47	0.81	0.19	[−0.02, 0.40]	0.08	
(d)	Context of music	−0.20	0.58	−0.04	[−0.28, 0.20]	0.74	
(e)	Familiarity	−0.49	0.32	−0.13	[−0.30, 0.04]	0.14	
(f)	Text length	0.24	0.45	0.06	[−0.15, 0.26]	0.59	
(g)	Linguistic demands	0.35	0.33	0.09	[−0.09, 0.27]	0.31	
(h)	Visuals	0.39	0.35	0.10	[−0.08, 0.27]	0.27	
							${R_{adj}}^2 = 0.71^{**}$

b represents unstandardized regression weights; beta indicates the standardized regression weights; LL and UL indicate the lower and upper limits of a confidence interval, respectively; R^2 represents the adjusted determination coefficient; and **indicates $p < 0.01$.

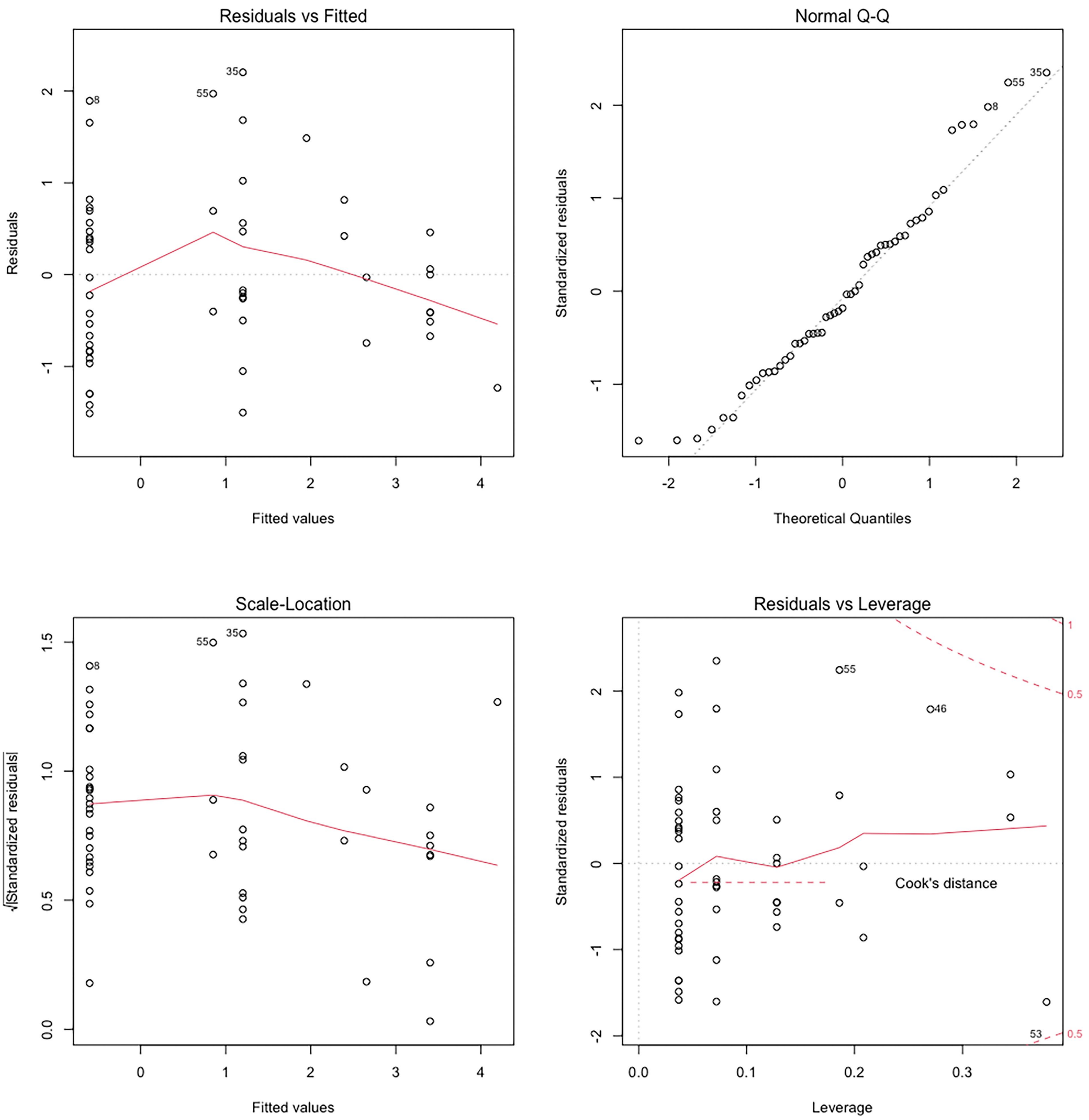


FIGURE 3
Four residual plots for regression Model 1 (cognitive requirements). The “Residuals vs. Fitted” plot (upper left) shows the fitted and unstandardized residual values. The “Normal Q-Q” plot illustrates the correct model specification and shows that the residuals are normally distributed. The assumption of homoscedasticity is shown in the “Scale-Location” diagram on the lower left. The outliers and influential values are shown on the plot “Residuals vs. Leverage” on the lower right (see also Luhmann, 2020, pp. 238–255; Field et al., 2012, pp. 266–276).

homoscedasticity assumption, and the identification of outliers and influential values.

Conclusion

Our findings show that the differences in item difficulty were predicted by the identified item characteristics. An important result was the categorization of the item characteristics into three domains: *cognitive requirements*, *knowledge*, and *formal item features*. A comparison of three regression models confirmed that Model 1, which held four predictors of the domain “cognitive requirements,” best fit the study data $R_{adj}{}^{2} = 0.71)$. The two regression models comprising predictors of the domains “knowledge” and “formal item features” failed to explain more variance. The strongest predictor in the final model was “reference to musical attributes” ($\beta = 0.46, \beta = 0.51)$, followed by “cross-sentence argumentation” ($\beta = 0.37)$ and “dialogical argumentation” ($\beta = 0.20)$.

An interesting finding was that items containing visuals and longer or linguistically more complex texts were not more difficult. This is especially surprising since item characteristics related to reading skills have usually been found to increase item difficulty (see "item difficulty and item characteristics"). Therefore, we do not claim that reading skills are generally irrelevant for the MARKO test. On the contrary, we assume that linguistic skills are *particularly* important, which is reflected in two of the characteristics of the cognitive domain. Our analyses show that these specific features (cross-sentence argumentation and dialogical argumentation) were more important than the length or grammatical structure of the reading text. Thus, for individuals who were able to use complex and dialogical argumentation, it seemed to make no difference whether or not they had to read a great deal of text before completing an item. Technically speaking, we argue that it can be assumed that the length and complexity of an item text are relevant, in principle, but presumably, the linguistic features are confounded with each other so that only the strongest or most difficult characteristics could eventually be used as predictors.

Furthermore, we assumed that items containing music that was familiar to the participants were easier, but the respective predictor was not significant ($\beta = -0.49, p = 0.12$). However, it is important to note that we had little information about which musical pieces the students were familiar with. Therefore, further research needs to investigate the possible relation between familiarity with a musical genre and item difficulty.

Our research findings also have important implications for the music classroom and the question of how music-related argumentative competence can be fostered. The strongest predictor "reference to musical attributes" suggests that music-related perception is highly relevant when engaging in music-related argumentation. Thus, before being able to name a specific musical attribute, it first has to be perceived (see also Koelsch, 2019 and "domain 1: cognitive requirements"). The predictors "cross-sentence" and "dialogical argumentation" were both related to linguistic competence, pointing to the importance of linguistic skills when engaging in music-related argumentation. Further research needs to examine the interrelation between linguistic skills and music-related argumentative competence.

Although our findings seem promising, there are also some limitations of our methodological approach to music-related argumentation. Argumentation is an interactive event and an exchange of arguments with a real opponent can only be represented to a limited extent in a competency test. Although there were several items in the final MARKO test that imitated dialogical situations (such as the item "Eurovision Song Contest"), a competency test can never be as interactive as a conversation with an 'actual' person.

Our findings about the relations between item characteristics and item difficulty contribute to a better understanding of music-related argumentative competence in general and the validity of the competence test in particular. Since the final regression model only consists of item characteristics based on central assumptions hypothesized in the theoretical MARKO competency model (Rolle, 2013), this can be interpreted as proof of the construct validity of the MARKO test according to Borsboom et al. (2004). More specifically, our analyses support our assumption that the item characteristics ("attributes" in Borsboom et al., 2004's terminology) not only exist, but variations in the item characteristics causally produce variation in the competency test outcome. Furthermore, they can provide valuable information for scale anchoring in future studies (Hartig et al., 2012). The identified item characteristics can be important in developing further items measuring music-related argumentative competence, making it possible to determine beforehand which tasks are easier or more difficult and, therefore, can be developed for a specific requirement.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: https://osf.io/t5cb8/?view_only=185c322c6dbf4a1d92aa88cd30b3afae.

Author contributions

All authors contributed to the design of the study and to the identification of the item characteristics of the study ("Identification and Categorization of Item Characteristics"). JE drafted the manuscript; JE and JK wrote sections of the manuscript. JE and JK conducted the statistical analyses; JE finalized statistical analyses. All authors revised this manuscript critically and made improvements on it. All authors approve the final version of the manuscript.

Funding

This project was supported by a grant from the Research Council of Norway (grant no. 320141).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Beck, B., and Klieme, E. (Eds.). (2007). *Sprachliche Kompetenzen–Konzepte und Messung*: DESI-Studie (Deutsch Englisch Schülerleistungen International). Beltz.

Borsboom, D., Mellenbergh, G. J., and van Heerden, J. (2004). The concept of validity. *Psychol. Rev.* 111, 1061–1071. doi: 10.1037/0033-295X.111.4.1061

Bossen, A. (2017). "Sprache als Gegenstand der Musikpädagogischen Forschung und des musikdidaktischen Diskurses im Kontext einer Sprachbildung im Fach," in *Sprache im Musikunterricht. Ausgewählte Aspekte Sprachbewussten Handelns im Kontext von Inklusion*. eds. A. Bossen and B. Jank (Potsdam: Universitätsverlag Potsdam), 21–54.

Canale, M., and Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Appl. Linguis.* 1, 1–47. doi: 10.1093/applin/1.1.1

Cohors-Fresenborg, E., Sjuts, J., and Sommer, N. (2004). "Komplexität von Denkvorgängen und Formalisierung von Wissen," in *Mathematische Kompetenzen von Schülerinnen und Schülern in Deutschland. Vertiefende Analysen im Rahmen von PISA 2000*. ed. M. Neubrand (Wiesbaden: VS Verlag), 109–144.

Eemeren, F. H. V., Garssen, B., Krabbe, E. C. W., Snoeck Henkemans, A. F., Verheij, B., and Wagemans, J. H. M. (2014). *Handbook of Argumentation Theory*. Dordrecht: Springer.

Ehninger, J. (2022). *Musikbezogenes Argumentieren. Testentwicklung und Kompetenzmodellierung* [Doctoral dissertation, University of Cologne]. KUPS. Available at: https://kups.ub.uni-koeln.de/61290/ (Accessed November 5, 2022).

Ehninger, J., Knigge, J., Schurig, M., and Rolle, C. (2021). A new measurement instrument for music-related argumentative competence: the MARKO competency test and competency model. *Front. Educ.* 6, 1–10. doi: 10.3389/feduc.2021.668538

Field, A., Miles, J., and Field, Z. (2012). *Discovering Statistics Using R*. Los Angeles: Sage Publications.

Hartig, J. (2007). "Skalierung und Definition von Kompetenzniveaus," in *Sprachliche Kompetenzen–Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)*. eds. B. Beck and E. Klieme (Weinheim: Beltz), 83–99.

Hartig, J. (2008). "Psychometric models for the assessment of competencies," in *Assessment of Competencies in Educational Settings*. eds. J. Hartig, E. Klieme and D. Leutner (Göttingen: Hogrefe & Huber), 69–90.

Hartig, J., Frey, A., Nold, G., and Klieme, E. (2012). An application of explanatory item response modeling for model-based proficiency scaling. *Educ. Psychol. Meas.* 72, 665–686. doi: 10.1177/0013164411430707

Hartig, J., and Jude, N. (2007). "Empirische Erfassung von Kompetenzen und psychometrische Kompetenzmodelle," in *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik. Eine Expertise im Auftrag des Bundesministeriums für Bildung und Forschung*. eds. J. Hartig and E. Klieme (Berlin: BMBF), 17–36.

Hartig, J., and Klieme, E. (2006). "Kompetenz und Kompetenzdiagnostik," in *Leistung und Leistungsdiagnostik*. ed. K. Schweizer (Heidelberg: Springer Medizin), 127–143.

Heller, V., and Morek, M. (2015). Academic discourse as situated practice: an introduction. *Linguist. Educ.* 31, 174–186. doi: 10.1016/j.linged.2014.01.008

Jordan, A.-K. (2014). "Empirische Validierung eines Kompetenzmodells für das Fach Musik–Teilkompetenz," in *Wahrnehmen und Kontextualisieren von Musik* (Münster: Waxmann)

Jordan, A.-K., Knigge, J., Lehmann, A. C., Niessen, A., and Lehmann-Wermser, A. (2012). Entwicklung und Validierung eines Kompetenzmodells im Fach Musik: Wahrnehmen und Kontextualisieren von Musik [development and validation of a competence model in music instruction. Perception and contextualization of music]. *Z. Pädagogik* 58, 500–521. doi: 10.25656/01:10392

Knigge, J. (2010). *Modellbasierte Entwicklung und Analyse von Testaufgaben zur Erfassung der Kompetenz "Musik Wahrnehmen und Kontextualisieren"* [Doctoral dissertation, Universität Bremen]. Available at: http://nbn-resolving.de/urn:nbn:de:gbv:46-diss000120066 (Accessed November 5, 2022).

Knörzer, L., Stark, R., Park, B., and Rolle, C. (2016). "I like reggae and bob Marley is already dead": an empirical study on music-related argumentation. *Psychol. Music* 44, 1158–1174. doi: 10.1177/0305735615614095

Koelsch, S. (2019). "Neural basis of music perception: melody, harmony, and timbre," in *The Oxford Handbook of Music and the Brain*. eds. M. H. Thaut and D. A. Hodges (Oxford Academic), 187–211.

Kuhn, D. (2005). *Education for Thinking*. Cambridge: Harvard University Press.

Kultusministerkonferenz. (2005). Beschlüsse der Kultusministerkonferenz: Einheitliche Prüfungsanforderungen in der Abiturprüfung Musik. Available at: http://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/1989/1989_12_01-EPA-Musik.pdf

Lazarou, D., Sutherland, R., and Erduran, S. (2016). Argumentation in science education as a systemic activity: an activity-theoretical perspective. *Int. J. Educ. Res.* 79, 150–166. doi: 10.1016/j.ijer.2016.07.008

Luhmann, M. (2020). *R für Einsteiger: Einführung in die Statistik-Software für die Sozialwissenschaften*. Weinheim: Beltz.

Morek, M., and Heller, V. (2012). Bildungssprache–Kommunikative, epistemische, soziale und interaktive Aspekte ihres Gebrauchs. *Z. Angew. Linguisik* 57, 67–101. doi: 10.1515/zfal-2012-0011

Morek, M., Heller, V., and Quasthoff, U. (2017). "Erklären und Argumentieren. Modellierungen und empirische Befunde zu Strukturen und Varianzen," in *Begründen–Erklären–Argumentieren*. eds. I. Meißner and E. L. Wyss (Tübingen: Stauffenburg), 11–46.

Morrison, S. J., and Demorest, S. M. (2009). Cultural constraints on music perception and cognition. *Prog. Brain Res.* 178, 67–77. doi: 10.1016/S0079-6123(09)17805-6

Nold, G., and Rossa, H. (2007). "Hörverstehen," in *Sprachliche Kompetenzen-Konzepte und Messung. DESI-Studie (Deutsch Englisch Schülerleistungen International)*. eds. B. Beck and E. Klieme (Weinheim: Beltz), 178–196.

Prenzel, M., Häußler, P., Rost, J., and Senkbeil, M. (2002). Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen? *Unterrichtswissenschaft* 30, 120–135. doi: 10.25656/01:7682

Quasthoff, U., Heller, V., and Morek, M. (2020a). "Diskurskompetenz und diskursive Partizipation als Schlüssel zur Teilhabe an Bildungsprozessen: Grundlegende Konzepte und Untersuchungslinien," in *Diskurserwerb in Familie, Peergroup und Unterricht: Passungen und Teilhabechancen*. eds. U. Quasthoff, V. Heller and M. Morek (Berlin: De Gruyter), 13–34.

Quasthoff, U., Wild, E., Domenech, M., Hollmann, J., Kluger, C., Krah, A., et al. (2020b). "Familiale Ressourcen für den Erwerb von Argumentationskompetenz," in *Diskurserwerb in Familie, Peergroup und Unterricht: Passungen und Teilhabechancen*. eds. U. Quasthoff, V. Heller and M. Morek (Berlin: De Gruyter), 79–106.

Rapanta, C. (2018). *Argumentation Strategies in the Classroom* . Wilmington, DE: Vernon Press.

Rolle, C. (2013). Argumentation skills in the music classroom: a quest for theory. In *Artistry*. eds. A. de Vugt and I. Malmberg Innsbruck: Helbling, 37–150.

Utdanningsdirektoratet. (2020). *Curriculum for Music (MUS01-02)*. Available at: https://www.udir.no/lk20/mus01-02?lang=eng (Accessed November 5, 2022).