frontiers in Education

ORIGINAL RESEARCH
published: 02 June 2021
doi: 10.3389/feduc.2021.668538

A New Measurement Instrument for Music-Related Argumentative Competence: The MARKO Competency Test and Competency Model

Julia Ehninger[1]*, *Jens Knigge*[2]*, *Michael Schurig*[3] *and Christian Rolle*[1]

[1]Institute for Music Education, University of Cologne, Cologne, Germany, [2]Department for Arts and Culture, Nord University, Levanger, Norway, [3]Faculty of Rehabilitation Sciences, TU Dortmund University, Dortmund, Germany

OPEN ACCESS

Edited by:
Shaljan Areepattamannil, Emirates College for Advanced Education, United Arab Emirates

Reviewed by:
María Isabel de Vicente-Yagüe Jara, University of Murcia, Spain
Daniel Müllensiefen, Goldsmiths University of London, United Kingdom

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

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Education

Received: *16 February 2021*
Accepted: *10 May 2021*
Published: *02 June 2021*

Citation:
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. 6:668538. doi: 10.3389/feduc.2021.668538

In this paper, we introduce the MARKO competency test and competency model, a new measurement instrument for music-related argumentative competence (MARKO: Musikbezogene ARgumentationsKOmpetenz; German for music-related argumentative competence). This competence, which plays an essential role in school curricula, refers to the ability to justify, and defend judgments about music. The two main goals of this study were 1) to design an assessment test for music-related argumentation that fulfills psychometric criteria and 2) to derive competency levels based on empirical data to describe the cognitive dispositions that are necessary when engaging in argumentation about music. Based on a theoretical framework, we developed a competency test to assess music-related argumentative competence. After two pretests (n = 391), we collected data from 440 students from grade nine to the university level. The final test consisted exclusively of open-ended items, which were rated with coding schemes that had been designed for each item. After ensuring inter-rater reliability, we composed an item pool that met psychometric criteria (e.g., local stochastic independence and item homogeneity) and represented content-related aspects in a meaningful way. Based on this item pool, we estimated a one-dimensional partial credit model. Following a standard-setting approach, four competency levels were derived from the empirical data. While individuals on the lowest competency level expressed their own opinions about the music by referring to salient musical attributes, participants on the highest level discussed different opinions on the music, and considered the social and cultural context of the music. The proficiency scores significantly varied between grades. Our findings empirically support some theoretical assumptions about music-related argumentation and challenge others.

Keywords: music-related argumentation, competency, assessment, music, reasoning, item response theory, empirical research, musical judgment

INTRODUCTION

Music-related argumentative competence is "the (learnable) ability to justify and defend aesthetic judgments about music in a comprehensive, plausible, and differentiated way" (Knörzer et al., 2016, p. 2). In our everyday lives and in educational contexts, we frequently talk about music, and give reasons for our opinions. After a concert, audience members might discuss whether they liked what they heard. On social media, comments are posted below music videos that might even lead to an interactive discussion about the video. At rehearsals, band members often talk about their musical progress. Argumentation also plays an integral role in music lessons at schools.

Students' argumentation skills are considered important for overall educational success (Kuhn, 2005), and music-related argumentation is crucial in music education. Talking about music is part of many music practices and language skills are helpful in enhancing musical learning processes. Thus, they are incorporated into the German music curricula, where argumentation plays an essential role, usually in the competency domain of "reflection" (e.g., Ministerium für Schule und Berufsbildung des Landes Schleswig-Holstein, 2015; Ministerium für Schule und Bildung des Landes Nordrhein-Westfalen, 2019).

Music-Related Argumentation and Research on Competencies

Theories on argumentation date back to antiquity. Toulmin (2003), one of the pioneers of modern argumentation theory, claimed that even though certain aspects of argumentation practice are field invariant, others vary from field to field. For example, a mathematician will have to deal with different "forums," "stakes," and "contextual details" (Toulmin, 1992, p. 9) when reasoning about a mathematical problem than a lawyer who appears in court.

Music-related argumentation is a form of aesthetic argumentation. Claims to validity in this field differ from, for instance, claims to validity in natural science. If a person claims after a concert that they did not like the way the conductor interpreted the piece, the judgment does not refer only to the concert itself but also to their own impression of the music. Aesthetic judgments are neither merely subjective (e.g., expressing personal preferences) nor objective (e.g., referring to musical characteristics) but also relational (Rolle, 1999). They refer to the relationship between the person making a claim and the aesthetic object (Kant, 1790/2007, § 1). Nevertheless, judgments about music can claim intersubjective validity. If someone is stating that a melody is lovely, they are articulating their aesthetic experience and may try to convince others by making the experience comprehensible for them (Knörzer et al., 2016, p. 2). Rolle (1999, p. 115) suggested that aesthetic judgments are recommendations. The judgments encourage others to perceive the aesthetic object in a certain way (see Stevenson, 1950).

Argumentation in music therefore needs a theoretical approach that takes into account interactive communication as well as field-dependent aspects. Rolle (2013) suggested a competency model for music-related argumentation that integrates theoretical assumptions on aesthetic argumentation (see above), general argumentation theories emphasizing the dialogical structure of argumentation (Eemeren et al., 2014, ch. 10; Wohlrapp, 2014), research on reflective judgment (King and Kitchener, 2004), and concepts from art education (Parsons, 1987). In his competency model, Rolle distinguished several levels of music-related argumentation. While people on lower levels refer only to the objective properties of the music, such as its musical attributes or expressive qualities, people on higher levels combine the former two aspects and are able to consider different aesthetic conventions or cultural practices (Rolle, 2013, p. 146). People on lower levels assume that different musical judgments are a matter of taste, whereas people on higher levels can reflect on their own musical preferences and integrate different perspectives and counterarguments into their reasoning.

Little research has been conducted in this field. Knörzer et al. (2016) carried out the first empirical study on Rolle's model. In their study, 37 participants listened to two versions of the same musical piece. They were then asked which of the two versions they liked better and why. Knörzer et al. divided their sample into three groups according to their expertise (high school students, university students majoring in music education, professional musicians, and music educators). The authors of the study analyzed which aspects of the music the participants referred to when giving reasons for their judgment. While participants with the lowest expertise referred to subjective aspects in their reasoning, participants with higher musical expertise more often took context-specific background knowledge into account. Gottschalk and Lehmann-Wermser (2013) investigated the music-related argumentative competence of ninth graders analyzing discussions in the music classroom.

Much empirical research has been conducted on the development of empirically validated competency models since international large scale assessments such as PISA or TIMMS started to use domain-specific competency models as theoretical frameworks (e.g., Leutner et al., 2017). In this context, the definition of competency is mainly based on the theoretical work of Weinert (2001) who suggested that competencies are "context-specific cognitive dispositions that are acquired and needed to successfully cope with certain situations or tasks in specific domains" (Koeppen et al., 2008, p. 62; see also Hartig et al., 2008). In the "specific domain" of music education, however, research on competencies has been scarce. Apart from the KoMus project, which investigated students' competency to perceive and contextualize music (Jordan and Knigge, 2010; Jordan et al., 2012), and the KOPRA-M project, which dealt with music performance competency (Hasselhorn and Lehmann, 2015), no empirical research has been conducted on music-related competency modelling (for an overview, see Hasselhorn and Knigge, in press). Based on Weinert's conceptual work and against the background of Knörzer et al.'s (2016, p. 2) suggestion, we define music-related argumentative competence as follows: Music-related argumentative competence is the context-specific cognitive disposition that is acquired and

needed to justify and defend aesthetic judgments about music in a comprehensive, plausible, and differentiated way.

Research Goal

Our study was designed to empirically investigate theoretical assumptions about music-related argumentation. How is music-related argumentative competence structured? Which aspects play a role when people are reasoning about music? Which characteristics contribute to an argument being better or worse than others? The empirical study is based on Rolle's theoretical framework on music-related argumentation (Rolle, 2013); see (*Music-Related Argumentation and Research on Competencies*). In line with our overall goal, our first aim was to develop a competency test for music-related argumentation based on theoretical assumptions about the nature of music-related argumentation. After ensuring the psychometric properties of the test (i.e., model fit and reliability), our second aim was to model competence levels based on empirical data to show the challenges faced by the participants when reasoning about music. To our knowledge, this is the first empirical research endeavor on competency levels in this field.

MATERIALS AND METHODS

In this section, we describe the test design, data collection, data analysis, and methodological procedure for the specification of the competency levels.

We conducted analyses using a partial credit model (PCM; Bond and Fox, 2015) from the item response theory (IRT) framework. In IRT, the probability of solving an item depends on the difficulty of the item and the ability of the person trying to solve it. This approach makes it possible to estimate personal ability values (i.e., weighted likelihood estimation [WLE]) and item difficulties (Thurstonian thresholds in PCM) on a common scale and draw conclusions about the underlying latent trait.

Test Design

For the MARKO competency test (Musikbezogene ARgumentationsKOmpetenz; German for music-related argumentative competence), over 60 test items were designed, and tested during two piloting phases in 2017 and 2018. We examined several German state and federal-level school curricula as well as schoolbooks. We also considered items that had been developed in the context of the KoMus project (Knigge, 2010; Jordan et al., 2012) and preliminary empirical research on music-related argumentation by Knörzer et al. (2016).

Middle school, high school, and university students participated in the 90-min piloting sessions (n = 391). Since the test was administered online, the testing sessions usually took place in the computer rooms of the cooperating institutions during regular music lessons. In the testing sessions, the participants wore headphones, and individually sat in front of computers while listening to music, and watching videos. During the test, they were asked to state and justify their judgment in a written statement. While the majority of items tested in the first pilot phase had to be discarded, the items in the second phase were constantly revised based on feedback from teachers, participants, and research fellows.

Twenty-five items were successfully incorporated into the main study. The final test consisted exclusively of open-ended items. These items were analyzed in terms of inter-rater reliability, item fit indices, and item discrimination. During the piloting phase, it became clear that these types of items were especially suited for measuring music-related argumentative competence because in closed items, arguments cannot be produced but only evaluated. The average processing time of the participants varied greatly because of the different amounts of text that they produced. Some students merely produced a sentence per item, while others wrote lengthy paragraphs to justify their music-related judgment. Therefore, in the introduction to the final version of the test, we told the participants that it was not important to complete all items, and asked them to take their time. We used a rotated test design to collect as much data as possible on all 25 items. The 25 items were split into three sets, and the sets appeared in a different order in the three final test booklets.

Each test session began with a short verbal introduction by the test supervisor. The online test also included an explanatory introduction that elaborated on the nature of music-related judgments and musical terminology.

Data Collection and Participants

The data collection for the main study took place in 2019 at nine public high schools during regularly scheduled music lessons, and two universities (with students majoring in music education programs) in the state of North Rhine-Westphalia, Germany (n = 440) (three high school students in the sample were visiting from other schools). Of the participants, 44.5% were female. About one third of the students were in grade nine (age: 14–15), 24.5% were in grade ten (age: 15–16), 28.6% in grade eleven (age: 16–17), 5.5% in grade twelve (age: 17–18), and 7.7% were university students. The mean age was 16 years (SD = 2.79) and the duration of the test was 90 min. In addition to the competency test, demographic data were collected (gender, age, family migration history, and language), as well as data on musical experience (i.e., whether participants received musical instrument lessons), and musical sophistication (Gold-MSI general musical sophistication; Müllensiefen et al., 2014). The three test booklets were distributed almost evenly among the participating students (booklet I: 33.6%, booklet II: 33.9%, booklet III: 32.5%). Since the participants were told that they did not have to respond to all the items but should take their time, some participants did not complete all test items (26.2% missing values).

Data Analysis

Sample Item

The test items were designed to measure different aspects of music-related argumentative competence. Some items aimed at assessing subjective perspectives, others were designed to determine how participants referred to musical attributes, and still others focused on the dialogical aspects of argumentation. For example, the participants were asked to comment on a

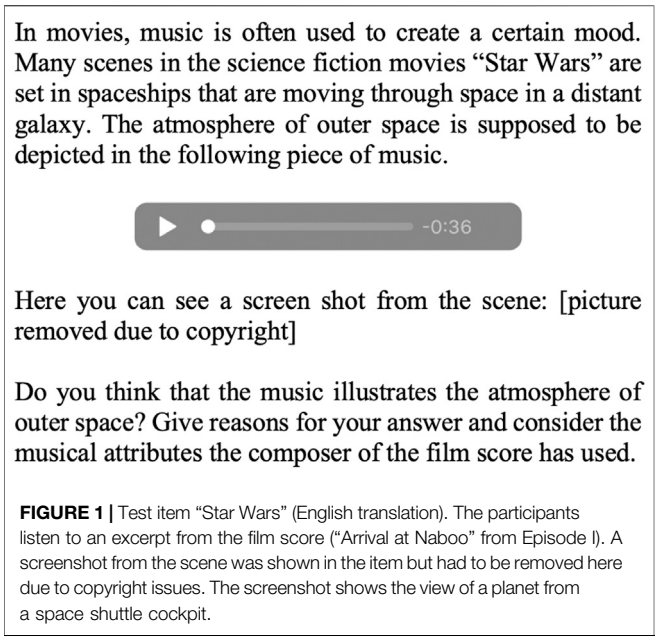

FIGURE 1 | Test item "Star Wars" (English translation). The participants listen to an excerpt from the film score ("Arrival at Naboo" from Episode I). A screenshot from the scene was shown in the item but had to be removed here due to copyright issues. The screenshot shows the view of a planet from a space shuttle cockpit.

discussion below a YouTube video, to react to a concert review in a newspaper, or to justify why they believed a song did or did not generate a certain atmosphere. Several incentives or triggers for possible argumentation were given in the test items. The following sample item exemplifies the items used in the test.

The sample item "Star Wars" was developed to assess how participants referred to musical attributes and to the generated musical atmosphere (**Figure 1**). In the sample item, there is an explicit request to refer to the musical attributes of the piece. In addition, the item text mentions the expressive qualities of the music and the desired mood that is supposed to be created ("atmosphere of outer space"). Thus, two types of references are suggested in the item, which also play a role in the theoretical model: musical attributes and expressive qualities of the music (Stages 3 and 4 in Rolle, 2013, p. 146). The supplementary material includes another sample item ("Eurovision Song Contest") focusing on the dialogical aspect of argumentation (see **Supplementary Material**, section 1).

As the test consisted exclusively of open-ended items, coding schemes had to be developed to rate the items.

Coding Process

Coding schemes were developed for each test item in a predominantly inductive and explorative process. By developing the coding schemes inductively, we were able to ensure the specifics of each item, since we also observed response behavior that could not be attributed to theoretical assumptions. **Table 1** gives an overview of the coding scheme used to rate test answers of the sample item "Star Wars." Test answers were rated with one point if the test taker referred only to the musical atmosphere or mentioned only salient (i.e., basic) musical attributes. Two points were assigned if a connection between musical attributes and the generated atmosphere was established. If participants went into further detail on specific aspects of the music, three points were given. Two raters coded approximately 15% of all the collected test data. The inter-rater reliability ranged from good to very good (Cohen's $\kappa = 0.73–0.94$ (calculation with linear weights)). The coding scheme for another item ("Eurovision Song Contest") is included in section 1 of the **Supplementary Material**.

Item Selection

The analyses were carried out with R version 3.6.2 (R Core Team, 2019) with the packages TAM (Robitzsch et al., 2020) and eRm (Mair et al., 2020). A one-dimensional partial credit model was estimated with the data we collected for the main study. Due to computational reasons, we conducted analyses for participants that had values for at least eight test items (pairwise deletion). 27 participants had to be excluded and analyses were conducted with $n = 440$. Missing values were not imputed.

We ensured that standards related to classical test theory were met (Wu et al., 2016, ch. 5). Item categories were collapsed if the relative frequency of the category was below 5%. We also monitored whether the item difficulty (i.e., Thurstonian thresholds) of the item categories appeared in the right order. As part of the criteria for Rasch conformity, Mean Squared Residual (MSQ) based item fit indices were calculated, considering conventional cut-off criteria (Ames and Penfield, 2015; Bond and Fox, 2015). In addition, item discrimination was determined as the point-biserial correlation of the item response category with the person ability (WLE) measured in the test. In a visual inspection, the expected item characteristic curves were compared with the empirically observed ones.

The global fit of the model and the assumption of local stochastic independence were examined with Q3 statistics, graphical model tests, and the Wald test. While 2.33% (Q3) and 0.67% (aQ3) of all 600 item pairs showed values above the cut-off criterion > 0.2 (Chen and Thissen, 1997), the mean of all Q3 and aQ3 values was close to zero (Q3: $M = -0.05$, $SD = 0.07$; aQ3: $M < 0.01$, $SD = 0.07$). Andersen's likelihood ratio test showed a significant result when the sample was split into two subsamples using a random split criterion (even vs. uneven case number) and gender as a split criterion (male vs. not male). Therefore, we also conducted graphical likelihood ratio model tests. No item categories were graphical outliers. The confidence ellipses intersected or were close to the identity line ($x = y$ line; **Figure 2**). Wald test results did not show anomalies, except for one item response category in the gender subsamples.

To ensure the fairness of the test, we conducted analyses for differential item functioning (DIF) with the group variables gender, language use at home, and with a variable specifying whether participants received musical instrument lessons. We followed the categorization proposed by the Educational Testing Service, assuming that an effect size ≥ 0.64 logits indicates moderate to large DIF (Trendtel et al., 2016, p. 131). Following this categorization, one item showed moderately significant DIF for participants who did not receive musical instrument lessons (−0.69 logits), and one item showed significant DIF for male participants (−0.72 logits). We kept both items in the item pool because the irregularities were not deemed detrimental.

TABLE 1 | Coding scheme for the sample item "Star Wars" (condensed and simplified version).

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 excited 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)

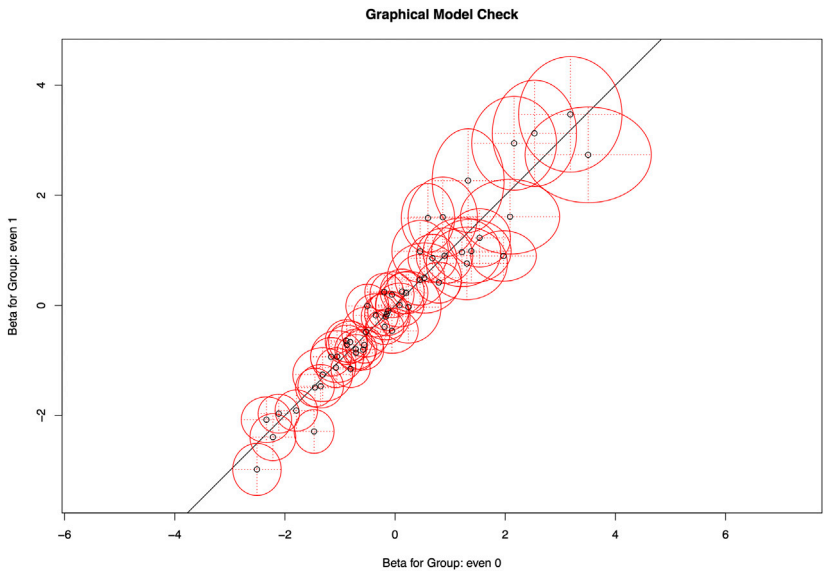

FIGURE 2 | Graphical model test (split criterion: even-uneven case number).

None of the remaining 25 test items were eliminated from the item pool that was used for the final computation of the model. However, eight item categories had to be collapsed due to misfitting item characteristics in terms of item difficulty.

Modeling Competency Levels

Conclusions about the nature of music-related argumentation can only be drawn through a content-related description of competency levels. Following this approach, the requirements that the participants must meet during the test can be determined. An important prerequisite for the criterion-related descriptions of the competency levels is the IRT scaling of the test.

In accordance with the bookmark method, which is a standard-setting procedure (e.g., Lewis et al., 2012), criteria-oriented competence levels were derived from the empirical data using external criteria from the theoretical model (Rolle, 2013) and the coding schemes. In the standard-setting procedure, item categories were ordered by their Thurstonian thresholds in an item-person map (Wright map) according to their 65% probability of solving the item response category correctly (with the R package WrightMap; Torres Irribarra and Freund, 2020). During the test design process, several skills and abilities that the participants had to master to solve the items were inspected. On the one hand, the coding schemes contained a lot of information about which competencies had to be mastered to solve a certain item response category. On the other hand, the test items had been designed based on Rolle's model and school curricula. These frameworks include assumptions about abilities

TABLE 2 | Information criteria for the estimated models. Model A represents a one-dimensional model. In model B, four items were assigned to a second dimension, and in model C, single-item categories were attributed to a second dimension.

Model	Loglike	Deviance	Npars	Nobs	AIC	BIC	AIC3	AICc	CAIC
Model A	–6,720.02	13,440.03	54	440	13,548.03	13,768.72	13,602.03	13,563.46	13,822.72
Model B	–6,641.70	13,283.39	129	440	13,541.39	14,068.59	13,670.39	13,649.58	14,197.59
Model C	–8,421.70	16,843.40	107	440	17,057.40	17,494.69	17,164.40	17,127.02	17,601.69

and task characteristics that can be crucial to the item-solving process (e.g., reference to salient vs. differentiated musical attributes, taking into account the expressive qualities of the music, and dealing with different perspectives on the musical piece).

Stemming from these a priori specified assumptions, three of the authors discussed which cognitive processes played a crucial role when solving an item. Every item response category in the ordered item booklet was reviewed and discussed in depth in terms of the relevant "knowledge, skills, and abilities" (Karantonis and Sireci, 2006, p. 5) that the participants had to master to solve the items. In this manner, cut scores were set for the competency levels; an accordingly qualified student is expected to have mastered the items below the cut score but not yet expected to have mastered the items above the bookmark. Mastery refers to having a 65% chance of solving the item. These cut scores were discussed and readjusted in an iterative process.

RESULTS

The Test

The final MARKO test consisted of 25 items (23 polytomous and 2 dichotomous items). The items that were selected showed appropriate infit (0.79–1.22) and outfit values (0.78–1.36; see **Supplementary Table S2**). Relative item frequency (percentage of participants who solved an item response category) ranged from 0.05 to 0.79. Thurstonian thresholds appeared in the right order. EAP/PV reliability equaled 0.91, and WLE reliability was 0.90 (see Adams, 2005; Rost, 2004, pp. 380–382 for an overview of test reliability measures).

The Model

In addition to a one-dimensional model (model A), we also estimated two two-dimensional models (**Table 2**). In model B, four items addressing the social context of the presented music were assigned to a second dimension. In model C, single item response categories focusing on the social context of the music were assigned to another dimension. In both models, the two dimensions had a correlation of $r = 0.96$. Exploratory factor analyses did not provide meaningful results due to low or double loadings. The analyses indicate that the item pool does not support multidimensionality. The resulting model is therefore a one-dimensional PCM.

Competency Levels and Proficiency Scores

Four competency levels were derived from our data following the standard-setting approach described in *Modeling Competency Levels*. The item categories were ordered by their item difficulty (i.e., 65% solution probability) in a Wright map (see **Figure 3**). The relevant abilities that the participants had to master to solve an item were identified. In this manner, conclusions about the competency levels were drawn.

Four competency levels were derived from the empirical data. Individuals on the lowest level, level A, express their own opinions about the music, and refer to salient musical attributes (e.g., "loud" and "fast") in their judgments. Participants on level B additionally report various opinions on the music. Whereas individuals on level B refer to several salient musical attributes, students on level C refer to musical attributes in detail in their judgments. Finally, individuals on level D discuss different opinions on the music and take into account the social and cultural context of the music (see **Table 3**).

While most ninth graders achieved competency level A, the majority of twelfth graders and university students obtained competency level C or D (**Figure 4**).

We estimated the participants' proficiency scores (person ability) using WLE (**Figure 5**). The numerical proficiency scores varied slightly but significantly between students in different grades, F (4, 434) = 44.85, $p < 0.01$, $\eta^2 = 0.29$. Post hoc analyses were conducted using pairwise t tests with pooled *SD*. With two exceptions (Grade 10 vs. Grade 11 and Grade 12 vs. university), the comparisons show significant results and medium to large effect sizes (see **Supplementary Table S3**). Female participants performed moderately better than male participants, t (418.58) = −4.21, $p < 0.01$, $\delta = 0.41$. Students who were taking musical instrument or voice lessons also had considerably better results, t (319.66) = −6.74, $p < 0.01$, $\delta = 0.68$. Participants who mostly spoke German at home performed significantly better as well, t (419.13) = −4.42, $p < 0.01$, $\delta = 0.42$. The general musical sophistication of the participants was assessed with a subscale from the Gold-MSI study (Müllensiefen et al., 2014) which consisted of 18 items (α =0.87). One item had to be removed from the scale due to low item-total correlation. The proficiency scores and general musical sophistication mean scores were significantly correlated, $r = 0.41$, $p < 0.01$.

DISCUSSION

The two main goals of this study were 1) to design an assessment test for music-related argumentation that fulfills psychometric criteria and 2) to model competency levels based on empirical data to describe the cognitive abilities that people must master when engaging in music-related argumentation. Competency modeling in this field of study is still in its infancy, and our project is one of the first empirical endeavors in this field.

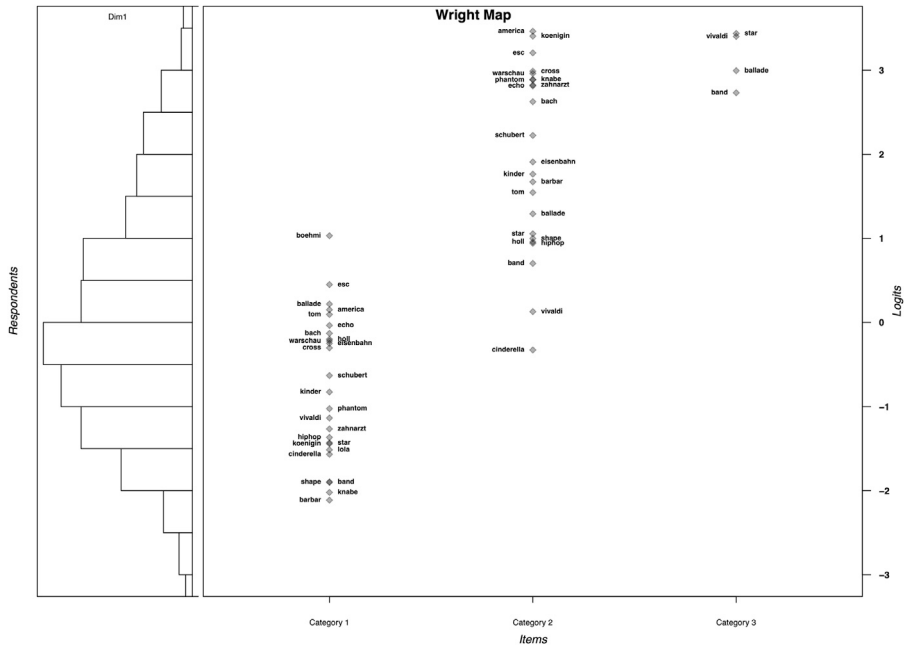

FIGURE 3 | Wright map (person-item map) with item categories ordered by 65% solution probability. On this map, the difficulty of the item response categories (Thurstonian thresholds) are placed on the same metric scale as the proficiency scores (person ability measures). The left side of the figure shows a histogram of the proficiency scores and the right side shows the difficulty of the item categories.

TABLE 3 | Description of the competency levels. The sample answers are taken from the two sample items, "Star Wars" (**Figure 1**) and "Eurovision Song Contest" (see section 1 in the **Supplementary Material**; solution probability 65%). The second column ("Logits") refers to the person ability scores (weighted likelihood estimation [WLE]). The table is supposed to be read from bottom to top.

Level	Logits	Description of competency levels	Sample answers
D	>2.22	Individuals have mastered competencies on levels A, B, and C and are able to • discuss different opinions on the music • refer to musical norms and genre conventions in their reasoning • take into account the social and cultural context of the music presented	"Women empowerment is a very current topic that is important. 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)
C	≤2.22	Individuals have mastered competencies on levels A and B and are able to • base music-related judgments on detailed references to musical attributes and link them to the expressive quality and function of the music • refer to basic knowledge of musical norms and genre conventions in their reasoning	"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)
B	≤0.45	Individuals have mastered competencies on level A and are able to • report various opinions on the music • base music-related judgments on several salient musical attributes (e.g., tempo, dynamics, intonation, and genre characteristics) and link them to the expressive quality and function of the music	"The singer addresses a very important and current topic: social equality. However, I think the point is not convincingly communicated. The lyrics are presented with humor and thus don't mean anything." (VP_142)
A	≤ −0.83	Individuals are able to • express their own opinions about the music • base music-related judgments on salient musical attributes (e.g., tempo, dynamics, intonation, and genre characteristics) • refer to the expressive quality and function of the music in their reasoning	"Yes, because of the atmosphere that exists in space. The composer presented this very well." (VP_661)

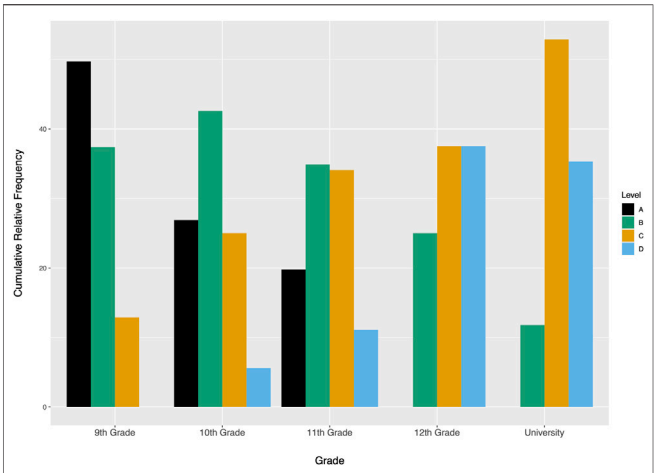

FIGURE 4 | Relative frequency of competency levels among grades (n = 439; Grade 9: n = 147; Grade 10: n = 108; Grade 11: n = 126; Grade 12: n = 24; University: n = 34).

The test items were developed considering the competency requirements in German school curricula and the theoretical assumptions about music-related argumentation by (Rolle, 2013). The test meets several psychometric criteria. Coding schemes of the exclusively open-ended test items suggest high inter-rater reliability. The empirical results show that the selected items represent a one-dimensional ability construct. EAP reliability amounted to 0.91, and global fit indices ranged from acceptable to good. Subgroup invariance was ensured with Q3 indices and graphical model inspection. From the empirical data, we were able to model competency levels following a standard-setting procedure. The resulting model describes the competencies that people show when giving reasons for their judgments about music.

We were able to model four competency levels that describe various aspects of music-related competence. While persons on the lowest level (A) are able to justify their judgments about music by referring to salient musical attributes or the overall atmosphere of the presented piece, persons on the highest level (D) are able to consider the social and cultural context of the music as well as the genre conventions.

Our findings empirically support some theoretical assumptions of Rolle (2013) theoretical competency model on music-related argumentation while challenging others. Whereas the theoretical model only talks in general terms about the ability to refer to the "objective" properties of music, the empirical study shows that the actual response behavior is more complex. The participants apparently find it easier to refer to the music in its entirety with the help of salient musical attributes (e.g., "loud" and "fast"). A differentiated reference to the presented music (e.g., to single passages in the music) occurs only on levels C and D. This phenomenon was described in the context of the KoMus project, which dealt with the competence to perceive and contextualize music (Jordan et al., 2012). In our empirical data, we also could not find the distinction made in the theoretical model, which assumes that on lower levels, a reference is made either to the objective properties of the music or to subjective impressions. At level A, for example, a salient musical attribute, such as "quiet," is used to describe the expressive quality of the music. We also found an even higher competency level, level E, as predicted by the theoretical model. Unfortunately, due to the low relative response frequency (<5%), the item categories had to be collapsed. Future studies could

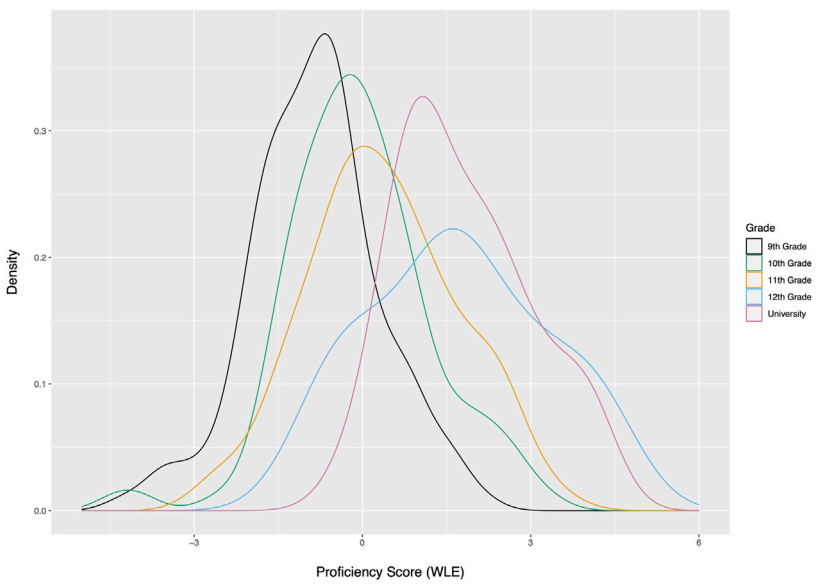

FIGURE 5 | Distribution of proficiency scores (weighted likelihood estimation) among grades as density plots.

include more university students to collect enough data on higher competency levels. In the theoretical model, (Rolle, 2013) assumed that people at a very low achievement level refer to authorities in their statements. This level was not found in our project (neither was it found during the pretest sessions or by Knörzer et al. (2016)).

While almost no ninth graders achieved competency level D, most of the twelfth graders and university students reached level C or D. In line with general expectations, students from higher grade levels performed significantly better on the test. Female participants performed better than the males. Students who took musical instrument lessons and participants who mostly spoke German at home performed significantly better as well Hasselhorn and Lehmann (2015) and Jordan (2014, pp. 141–142) also showed in their studies on music-related competencies that female participants and students who took musical instrument lessons performed significantly better on the music competency test. While the latter finding is not surprising and suggests that students who have acquired skills on a musical instrument perform better on music-related assessment tests, more research has to be conducted on the relationship between test performance and gender. Preliminary path analyses of our data show only a small effect of gender on proficiency scores when the variable "musical instrumental lessons" is controlled for. Gender-specific aspects in the music classroom have hardly been researched, but Heß (2018) as well as Fiedler and Hasselhorn (2020) showed that girls have a higher musical self-concept than boys.

Although our findings are promising, our study has some limitations. As mentioned earlier, the participants' processing time varied greatly. More competent students tended to write longer statements and therefore did not process as many items as less competent students did. Hence, the sample of this study contains systematic missing values that are correlated with the participants' proficiency scores ($r = -0.51$, $p < 0.01$).

In his theoretical model, Rolle (2013) took into account that argumentation is an interactive event and theorized about how individuals deal with counterarguments presented by an opponent. Though we designed several items imitating dialogical situations (e.g., the item "Eurovision Song Contest", see section 1 in the **Supplementary Material**), an assessment test will never be as interactive as a real conversation with an actual person. Interactive research settings, such as group discussions, could provide information about interactive verbal exchanges (see also Ehninger, 2021, for the impact of research settings and methodology on research on music-related argumentative competence).

The construct validity of the test instrument needs to be examined in future studies. General language skills likely play an essential role when reasoning about music. This interrelation should be explored in more detail, not least to assess the discriminant validity of the MARKO test. Future studies could develop a shortened version of the MARKO test that leaves extra time to assess participants' language skills. With this approach, one could analyze the interaction of domain-specific and overall language competencies. Research has shown that students' argumentative and linguistic skills are crucial to their domain-specific learning and overall educational success (Morek et al., 2017). Furthermore, it remains uncertain which other music-related factors influence participants' test results. In this study, we showed that participants who took musical instrument lessons performed significantly better on the test. Analyses showed a correlation between general musical sophistication and proficiency scores. It remains open whether musical preferences affected the test performance of the participants. In our study, musical preferences were assessed via a selection of six audio excerpts from six test items. The participants listened to these audio excerpts at the end of the session. However, we did not collect enough data to draw conclusions about the relationship between music-related argumentation and musical preferences.

Our study yielded important findings in a field that has been little researched. The one-dimensional model derived from empirical data allows a detailed description of four competency levels. On the basis of these levels, conclusions can be drawn about the attainment of music-related argumentative competence. The MARKO competency test and model can thus help enhance the understanding of learning processes and improve the assessment of music-related argumentative competence.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in the following online repository: https://osf.io/7tm9x/?view_only=87043a1db7c942dc9a8851af25025026

ETHICS STATEMENT

The data for the study was collected in accordance with the guidelines of the state of North Rhine-Westphalia, Germany [Schulgesetz für das Land Nordrhein-Westfalen (SchulG)]: BASS § 120 Abs. 4 SchulG. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

JE, JK, and CR contributed to the conception and design of the study. Test items for the first piloting study were designed by JE, JK, CR, and university students who also collected data for the first pilot study. JE revised and designed new items for the second pilot study supervised by JK and CR. JE collected the data for the second piloting phase and the main study and performed statistical analyses. MS and JK supported the analyses. JE drafted the manuscript. All other authors revised this manuscript critically and made improvements on it. All authors approve the final version of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2021.668538/full#supplementary-material

REFERENCES

Adams, R. J. (2005). Reliability as a Measurement Design Effect. *Stud. Educ. Eval.* 31 (2), 162–172. doi:10.1016/j.stueduc.2005.05.008

Ames, A. J., and Penfield, R. D. (2015). An NCME Instructional Module on Item-Fit Statistics for Item Response Theory Models. *Educ. Meas. Issues Pract.* 34 (3), 39–48. doi:10.1111/emip.12067

Bond, T. G., and Fox, C. M. (2015). *Applying the Rasch Model. Fundamental Measurement in the Human Sciences*. London: Routledge

Chen, W.-H., and Thissen, D. (1997). Local Dependence Indexes for Item Pairs Using Item Response Theory. *J. Educ. Behav. Stat.* 22 (3), 265–289. doi:10.2307/1165285

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. (2021). Wie lässt sich musikbezogene Argumentationskompetenz empirisch untersuchen? Über die empirische Erforschung einer facettenreichen Kompetenz. *Beiträge empirischer Musikpädagogik*. 12, 1–31. Available at: https://b-em.info/index.php/ojs/article/view/192

Fiedler, D., and Hasselhorn, J. (2020). Zum Zusammenhang von musikalischem Selbstkonzept und Motivation im Musikunterricht. *Beiträge empirischer Musikpädagogik*. 11, 1–34. Available at: https://b-em.info/index.php/ojs/article/view/187.

Gottschalk, T., and Lehmann-Wermser, A. (2013). "Iteratives Forschen am Beispiel der Förderung musikalisch-ästhetischer Diskursfähigkeit," in *Der lange Weg zum Unterrichtsdesign. Zur Begründung und Umsetzung fachdidaktischer Forschungs- und Entwicklungsprogramme*. Editors M. Komorek and S. Prediger (Münster: Waxmann), 63–78.

Hasselhorn, J., and Knigge, J. (in press). *"Technology-Based Competency Assessment in Music Education: The KOPRA-M and KoMus Tests," in Testing and Feedback in Music Education – Symposium Hannover 2017*) Editors A. Lehmann-Wermser and A. Breiter (Hannover: ifmpf).

Hasselhorn, J., and Lehmann, A. C. (2015). *"Leistungsheterogenität im Musikunterricht. Eine empirische Untersuchung zu Leistungsunterschieden im Bereich der Musikpraxis in Jahrgangsstufe 9," in Theoretische Rahmung und Theoriebildung in der musikpädagogischen Forschung*. Editors A. Niessen and J. Knigge (Münster: Waxmann), 163–176.

Heß, F. (2018). *Gendersensibler Musikunterricht. Empirische Studien und didaktische Konsequenzen*. Wiesbaden: Springer. doi:10.1007/978-3-658-19166-5

Hartig, J., Klieme, E., and Leutner, D. Eds. (2008). *Assessment of Competencies in Educational Settings*. Göttingen: Hogrefe.

Jordan, A.-K. (2014). *Empirische Validierung eines Kompetenzmodells für das Fach Musik – Teilkompetenz, 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. *Z. für Pädagogik.* 58 (4), 500–521.

Jordan, A.-K., and Knigge, J. (2010). "The Development of Competency Models: An IRT-Based Approach to Competency Assessment in General Music Education," in *The Practice of Assessment in Music Education: Frameworks, Models, and Designs*. Editor T. S. Brophy (Chicago: GIA), 67–86.

Kant, I. (1790/2007). *Critique of Judgment. Translated by J. C. Meredith*. Oxford: Oxford University Press.

Karantonis, A., and Sireci, S. G. (2006). The Bookmark Standard-Setting Method: a Literature Review. *Educ. Meas.* 25, 4–12. doi:10.1111/j.1745-3992.2006.00047.x

King, P. M., and Kitchener, K. S. (2004). Reflective Judgment: Theory and Research on the Development of Epistemic Assumptions through Adulthood. *Educ. Psychol.* 39, 5–18. doi:10.1207/s15326985ep3901_2

Knigge, J. (2010). *Modellbasierte Entwicklung und Analyse von Testaufgaben zur Erfassung der Kompetenz "Musik wahrnehmen und kontextualisieren" [dissertation]*. Universität Bremen. Available at: https://media.suub.uni-bremen.de/handle/elib/2844.

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 (5), 1158–1174. doi:10.1177/0305735615614095

Koeppen, K., Hartig, J., Klieme, E., and Leutner, D. (2008). Current Issues in Competence Modeling and Assessment. *Z. für Psychol./J. Psychol.* 216, 61–73. doi:10.1027/0044-3409.216.2.61

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

Leutner, D., Fleischer, J., Grünkorn, J., and Klieme, E. (2017). *Competence Assessment in Education*. Basel: Springer. doi:10.1007/978-3-319-50030-0

Lewis, D. M., Mitzel, H. C., Mercado, R. L., Patz, R. J., and Schulz, E. M. (2012). "The Bookmark Standard Setting Procedure," in *Setting Performance Standards: Foundations, Methods, and Innovations*. Editor G.J. Cizek. 2nd ed. (New York: Routledge), 225–253.

Mair, P., Hatzinger, R., Maier, M. J., Rusch, T., and Debelak, R. (2020). *eRm: Extended Rasch Modeling*. R package version 1.0-1. Available at: https://CRAN.R-project.org/package=eRm.

Ministerium für Schule und Berufsbildung des Landes Schleswig-Holstein (2015). *Fachanforderungen Musik. Allgemeinbildende Schulen. Sekundarstufe I. Sekundarstufe II*. Kiel: Ministerium für Schule und Berufsbildung des Landes Schleswig-Holstein.

Ministerium für Schule und Bildung des Landes Nordrhein-Westfalen (2019). *Musik. Kernlehrplan für das Gymnasium Sekundarstufe I in Nordrhein-Westfalen*. Düsseldorf: Ministerium für Schule und Bildung des Landes Nordrhein-Westfalen.

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

Müllensiefen, D., Gingras, B., Musil, J., and Stewart, L. (2014). The Musicality of Non-musicians: an index for Assessing Musical Sophistication in the General Population. *PLoS ONE.* 9 (2), e89642. doi:10.1371/journal.pone.0089642

Parsons, M. J. (1987). *How We Understand Art. A Cognitive Developmental Account of Aesthetic Experience*. Cambridge: University Press.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R version 3.6.2. Vienna: R Foundation for Statistical Computing.

Robitzsch, A., Kiefer, T., and Wu, M. (2020). TAM: Test Analysis Modules. *R Package Version 3.5-19*. Available at: https://CRAN.R-project.org/package=TAM.

Rolle, C. (1999). *Musikalisch-ästhetische Bildung. Über die Bedeutung ästhetischer Erfahrung für musikalische Bildungsprozesse*. Kassel: Bosse.

Rolle, C. (2013). Argumentation Skills in the Music Classroom: A Quest for Theory. In: *European Perspectives on Music Education 2: Artistry*. Editors A. de Vugt and I. Malmberg (Innsbruck: Helbling), 137–150.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion*. Bern: Huber. doi:10.1007/978-3-322-80662-8

Stevenson, C. L. (1950). *"Interpretation and Evaluation in Aesthetics," in Philosophical Analysis*. Editor M. Black (Ithaca: Cornell University Press), 341–383.

Torres Irribarra, D., and Freund, R. (2020). *Wright Map: IRT Item-Person Map with ConQuest Integration*. R package version 1.2.3.

Toulmin, S. E. (1992). "Logic, Rhetoric and Reason. Redressing the Balance," in *Argumentation Illuminated*. Editors F. H. V. Eemeren, R. Grootendorst, J. A. Blair, and C. A. Willard (Amsterdam: Sicsat), 3–11.

Toulmin, S. E. (2003). *The Uses of Argument*. Cambridge: Cambridge University Press. doi:10.1017/cbo9780511840005

Trendtel, M., Schwabe, F., and Fellinger, R. (2016). "Differenzielles Itemfunktionieren in Subgruppen," in *Large-Scale Assessment mit R. Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung*. Editors S. Breit and C. Schreiner (Wien: Facultas), 111–147.

Weinert, F. E. (2001). "Concept of Competence: A Conceptual Clarification" in *Defining and Selecting Key Competencies*. Editors D.S. Rychen and L.H. Salganik (Göttingen: Hogrefe), 45–65.

Wohlrapp, H. (2014). *The Concept of Argument*. Dordrecht: Springer.

Wu, M., Tam, H. P., and Jen, T.-H. (2016). *Educational Measurement for Applied Researchers. Theory into Practice*. Singapore: Springer. doi:10.1007/978-981-10-3302-5

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.