frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 16 August 2024
DOI 10.3389/feduc.2024.1332821

OPEN ACCESS

EDITED BY
Francesca Giovanna Maria Gastaldi,
University of Turin, Italy

REVIEWED BY
Elena Mirela Samfira,
University of Life Sciences "King Mihai I" from Timisoara, Romania
Ting Wang,
American Board of Family Medicine, United States

*CORRESPONDENCE
Kirsti Nørstebø
kirsti.norstebo@nord.no

RECEIVED 03 November 2023
ACCEPTED 05 August 2024
PUBLISHED 16 August 2024

CITATION
Nørstebø K and Knigge J (2024) KoMus and KOPRA-M: psychometric analysis of two musical competency tests adapted for Norwegian primary school students.
Front. Educ. 9:1332821.
doi: 10.3389/feduc.2024.1332821

KoMus and KOPRA-M: psychometric analysis of two musical competency tests adapted for Norwegian primary school students

Kirsti Nørstebø* and Jens Knigge

Department for Arts and Culture, Nord University, Bodø, Norway

This paper presents Norwegian adaptations of the KoMus and KOPRA-M assessments designed to evaluate music-related competencies. Our research delves into the assessments' alignment with the curriculum and investigates the psychometric properties of the Norwegian versions using a sample of Norwegian fifth graders (KoMus: $N = 374$, KOPRA-M: $N = 370$). Furthermore, it provides practical illustrations of how these item response theory (IRT)–based assessments can be employed, including competency-level examples. The Norwegian short versions of the KoMus and KOPRA-M tests demonstrated robust psychometric characteristics (especially in terms of reliability, model, and item fit), making them promising instruments for the deeper exploration of musical competency within primary and lower secondary school contexts.

KEYWORDS

assessment and education, competency testing, music education, music-related competency, empirical research, quantitative research

1 Introduction

In Norwegian primary and lower secondary schools, music is a compulsory minor subject with its own curriculum (Kunnskapsdepartementet, 2019). This curriculum consists of several competence aims that cover the core areas of musical competence. However, we know very little about the learning outcomes, such as musical competencies, of music lessons in Norwegian primary and lower secondary schools, since empirical research in this area is scarce. Notably, there is a lack of quantitative assessment instruments for measuring students' competencies.

This absence of research means that there is a limited understanding of students' competencies, their developmental progress, and the factors influencing their development. Consequently, a crucial aspect of music education—and, by extension, music teacher education—lacks an evidence base. Therefore, the first essential step is to develop measurement tools that facilitate evidence-based professional and educational development.

To address this gap, this study proposes the adaptation of two internationally recognized competency tests—namely, the KoMus and KOPRA-M assessments—for the Norwegian context. In the following sections, we present (a) why we consider these two particular tests (originally in German) to be suitable for the Norwegian context, (b) what adaptations were necessary for Norwegian primary school students, and (c) what psychometric characteristics

emerge for the Norwegian versions based on a sample of Year 5 students ($N = 370$ for the KOPRA-M test and $N = 376$ for the KoMus test).

There were two main reasons for choosing fifth graders as the sample. First, this primarily methodological paper was produced within the context of a larger quasi-experimental study (OutMus[1]) that deals with the longitudinal competence development of fifth graders; therefore, measurement instruments that are adequate and sensitive for exactly this age group are needed. Second, the focus on fifth graders is reasonable, given that at this stage, all pupils in Norway continue to take lessons in music—a subject that becomes optional in later secondary grades.

2 Research on musical competencies

2.1 The concept of competency

In our study, we are particularly interested in competencies as defined by the curriculum. Curriculum development in Norway—as in many other countries—has been strongly influenced by international large-scale studies, such as the PISA studies conducted by the OECD. Therefore, the Norwegian primary school curriculum's focus on competence is not surprising because it mirrors the work that the psychologist Franz E. Weinert carried out for the OECD more than two decades ago (e.g., Weinert, 1999, 2001). According to Weinert (2001), competency is domain specific and context related. Klieme and Leutner (2006) built on this understanding by describing competency as encompassing the cognitive dispositions acquired by learning that are needed to handle given situations and solve specific tasks. The definition of competence as context specific is made in clear distinction from intelligence (defined as a general cognitive ability; Hartig and Klieme, 2006, p. 129). In other words, competence is linked to the successful mastery of specific requirements—namely, subject-related requirements (mathematics, music, etc.). In psychological research, it is assumed that this domain specificity of competence goes hand in hand with its learnability. Whereas intelligence is regarded as a personal characteristic that is relatively stable over time and largely determined by biological factors, competencies are learnable because they are acquired through experience in specific situations (Hartig and Klieme, 2006, pp. 130–131). This psychological distinction between intelligence and competence corresponds to a certain extent to the music–psychological distinction between musical aptitude/ability and musical achievement/competence, as traced in the history of music test development over the last 100 years.

2.2 Literature review

In international music education and psychology (especially in Australia, England, and the United States), there is a long tradition of standardized music tests that goes back to the middle of the nineteenth century (see, e.g., Boyle and Radocy, 1987). This field continues to evolve, with recent advances allowing for the assessment of musical ability via online platforms in very short test durations (Correia et al., 2022; Strauss et al., 2024). The term "music test" is usually used to refer to a variety of measurement procedures. Such tests can be systematized into the following areas: musical aptitude and ability, musical achievement, musical performance, and musical attitude and appreciation (Boyle and Radocy, 1987). Only achievement tests are of interest in the present context. Although some of the content of aptitude tests overlaps with that of achievement tests, the two types of tests refer to clearly different theoretical constructs; that is, "aptitude or musicality tests aim to measure the innate potential of musical abilities (aptitude) independent of learning experiences. Musical achievement tests refer to the testing of musical skills learned through instruction" (Gembris, 1998, p. 111). The relationship between aptitude and achievement tests is thus comparable to the already discussed relationship between intelligence and competence. Therefore, the following considerations focus exclusively on musical achievement tests, which make up only a marginal share of the totality of the music tests. In the following sections, we briefly discuss the most relevant competency tests for our context (fifth graders in a Scandinavian school system) over the last 20 years (a more extensive review of current international test developments can be found in Brophy, 2019).

2.2.1 National assessments in the USA and Finland

In the USA, music has occasionally been included in the national assessments of educational progress (NAEP)—first in 1971 and most recently in 2016. The assessments for eighth graders were conducted according to a music assessment framework that covers "areas of process" (creating, performing, and responding) and "areas of content" (knowledge and skills). These tests were administered as both paper-and-pencil tasks and practical tasks wherein music had to be performed in front of judges (National Assessment Governing Board, 2016).

The reliability of the tests was evaluated in different ways for paper-and-pencil tasks and performance tasks. The results from the NAEP 1997 Arts Technical Analysis Report (Allen et al., 2004) showed adequate reliability, with alphas ranging from 0.63 to 0.77 for the different blocks of the music assessment. Alphas in this range are considered, according to COTAN (2019), to have sufficient reliability for conducting group-level analyses. Criticism of the NAEP in music has been directed at the validity of the test (e.g., Colwell, 1999).

Furthermore, in Finland, music has been part of national assessments since 2011. These tests are similar to the NAEP tests, including both paper-and-pencil tasks and performative tasks, and they are administered to ninth graders (Juntunen, 2017). Cronbach's alpha was used to assess their internal reliability, which was considered good (0.81 for the paper-and-pencil tasks and 0.92 for the production tasks) (Juntunen, 2017).

Juntunen (2017) highlighted issues with both the structure of music education in Finland and the design of the tests themselves. In Finland, the music curriculum is very open, granting teachers high autonomy, which leads to inconsistencies in the content taught across schools. The lack of standardized requirements further compounds this effect. Juntunen (2017) also criticized the assessment format, noting that the pencil-and-paper tasks inadequately evaluated only a fraction of musical competencies.

2.2.2 The KoMus test for measuring the competency of perceiving and contextualizing music

Jordan et al. (2012; see also Hasselhorn and Knigge, 2021) developed an item response theory (IRT)–based competency model

1 Further information is available on the project's homepage: https://site.nord.no/outmus/.

and associated test for the competency domain of "perceiving and contextualizing music" (KoMus) for sixth graders in lower secondary schools in Germany. The competency model and the associated test consist of the main dimension of perception and musical memory (D1) and three subdimensions: perception-based use of musical terminology (D2); musical notation (D3); and historical/cultural context knowledge (D4). The test is implemented as a technology-based competency assessment (TBCA; Hasselhorn and Knigge, 2021), which means that students complete the test individually in a browser via a computer or tablet (using headphones). The test can be distributed via different learning management or questionnaire platforms (e.g., Moodle or SoSci Survey). Complete test material is available upon request from the test developers.

The original KoMus test, published in 2012, consisted of 79 items and showed satisfactory to good reliability in all four dimensions (EAP/PV reliability: D1 = 0.82, D2 = 0.81, D3 = 0.79, D4 = 0.69). The test has been used in several studies in Germany and, even its short version (with only 29 items), has demonstrated consistently high psychometrical quality (e.g., Harnischmacher and Knigge, 2017; see also Hasselhorn and Knigge, 2021).

2.2.3 The KOPRA-M test for measuring music performance competency

The three-dimensional model and TBCA of music performance competency (KOPRA-M) were developed by Hasselhorn (2015; see also Hasselhorn and Knigge, 2021) for ninth graders in lower secondary schools in Germany. The dimensions of *singing*, *playing an instrument*, and *playing rhythms* were implemented in an Android app in which the participants can sing, perform rhythms, and perform melodies using a "colored music grid" (see also Figure 1) as their "instrument" (Hasselhorn, 2015).

The IRT-based KOPRA-M test, published in 2015, consists of 48 items and has shown very good reliability in all three dimensions (EAP/PV reliability: D1 = 0.96, D2 = 0.91, D3 = 0.92). The KOPRA-M test has been used in several studies in Germany and has consistently shown high psychometrical quality, even when used with younger and older (fifth-to tenth-grade) students (Lill et al., 2019). Complete material is available upon request from the test developers.

2.2.4 The MARKO test for measuring music-related argumentation competency

The music-related argumentation competency test (MARKO) was developed by Ehninger et al. (2021) and was designed to evaluate the ability to justify and defend judgments about music, an area of competence that is an integral part of many schools' music curricula. Based on a sample of participants ranging from ninth grade to university level, different levels of competence can be derived. The IRT-based test consists of 25 items, and both the EAP/PV reliability (0.91) and the WLE reliability (0.90) have shown high psychometric quality. Complete test material is available upon request from the authors.

2.2.5 Summary

For the purpose of our study, the music assessments provided by the NAEP, the Finnish national assessments, and the MARKO test did not align with the age range of the participants. Additionally, both the NAEP and the Finnish assessments rely on traditional pencil-and-paper formats, with performative aspects requiring face-to-face interactions with judges. Due to these factors, we chose to exclude these assessments from the current study. However, we identified the KoMus and KOPRA-M tests as fitting options for our targeted age group. Notably, these tests were developed using a state-of-the-art construction procedure (i.e., IRT-based analysis) to ensure the psychometric quality and implementation of technology-based versions. In this way, data collection and analysis are facilitated, hence promoting the objectivity, reliability, and validity of the tests (for a detailed discussion of the potential advantages and the psychometric effects of computer-based assessments, see Buerger et al., 2016, and Jurecka, 2008; for a music-specific discussion, see Hasselhorn and Knigge, 2021). For these reasons, we decided to use these tests for our research. What follows is a description of the Norwegian music curriculum and how it relates to the selected tests.

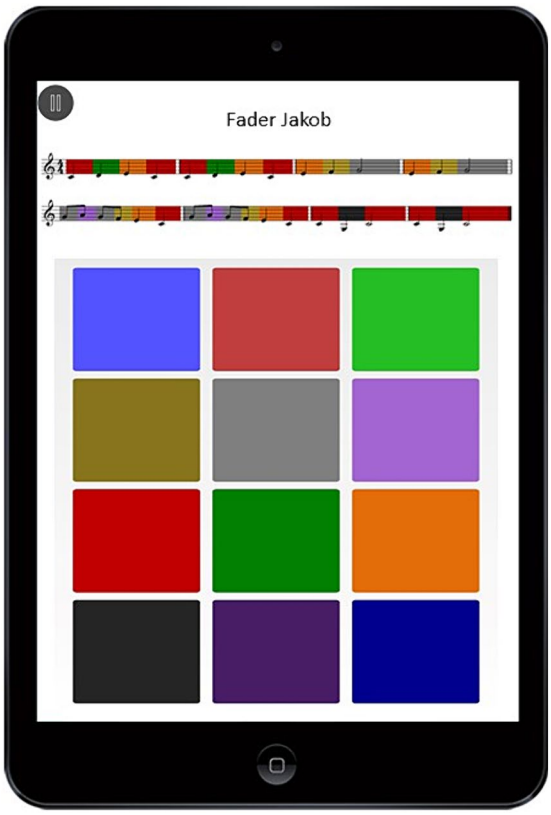

FIGURE 1
Item example from the KOPRA-M test (dimension: instrumental performance).

2.3 Comparison of the selected tests with the Norwegian primary school music curriculum

In Norway, the music curriculum (Kunnskapsdepartementet, 2019) describes how the subject is built around its relevance and central values, interdisciplinary topics, and basic skills.

In the relevance and central values section of the curriculum, it is emphasized that "the subject should promote the enjoyment of music and give a sense of mastery, and students should experience that their voice is important in the shared environment with their peers" (Kunnskapsdepartementet, 2019). Furthermore, the basic skills

section elaborates, for example, on the development of oral skills in music: "The development of oral skills progresses from being able to talk about one's own experiences and using simple techniques to being able to describe more complex music-related topics, aesthetic perceptions, musical techniques and the functions of music in more detail" (Kunnskapsdepartementet, 2019).

Four core areas (experiencing music, making music, performing music, and cultural understanding) describe the academic content of the music subject, in addition to the competence goals and formative assessment described after grades 2, 4, and 7. In the following section, we examine the core areas covered by the KoMus and KOPRA-M tests.

Experiencing music is about perceiving music in different ways, with the goal of students developing a reflective relationship with music. This core area is covered by the first three dimensions of the KoMus test. Here, we find items that focus on perception, terminology, and different types of notations. Notation is mentioned in the subject's basic skills section, wherein writing and reading in music include graphic notation, written music, or written figures. A competency goal related to this core area is that the pupil should be able to "use subject-related terms in the description of and reflection on work processes, results, musical expressions and techniques" and "explore and present musical experiences and perceptions" (Kunnskapsdepartementet, 2019).

Performing music is about playing, singing, and dancing as an active participant. Through various creative processes, pupils are expected to participate and practice through crafts, interaction, performance, expression, and dissemination in different expressions and genres. A competency aim for this core area after grade 7 states that the pupil is "expected to perform a repertoire of music, songs, other vocal expressions and dance from the contemporary period and from past times" (Kunnskapsdepartementet, 2019). This area is in line with the content of the KOPRA-M test, which evaluates competencies in singing and playing instruments.

Cultural understanding is thematically about connecting music and society. The core element has a clear societal perspective on which a two-sided relationship between musical expression and society is based. A competency aim for this core area concerns how a pupil should be able to "reflect on how music can play different roles in developing the identity of individuals and groups" (Kunnskapsdepartementet, 2019). The fourth dimension of the KoMus test is relevant to this area of competence, encompassing items centered on historical and cultural contextual knowledge.

Based on the currently available international competency tests (see above), and in alignment with the Norwegian curriculum for music in primary school, we decided to adapt both the KoMus (Jordan et al., 2012) and the KOPRA-M (Hasselhorn, 2015) tests for the Norwegian context, since both have the greatest overlap in terms of content, they are available in versions for children in the relevant age range, they show high psychometric quality, and, last but not least, relevant materials for both tests are fully accessible via the test developers.

3 Materials and methods

Both tests (KoMus and KOPRA-M) were translated and adapted to suit the Norwegian context. The translations were performed by two Norwegian native speakers, one of whom was proficient in German. To ensure the accuracy of the content and meaning of the items, the items were reviewed by an expert in music test development who was a native German speaker fluent in Norwegian.

Pretests were carried out at three schools in spring 2022. The aim was to identify items that the participants found difficult to understand or difficult to read. The phrasing of some items had to be changed after pretesting. The pretests were also used to check the technical setup, since both tests were implemented in a web-based assessment framework (described in the following section).

3.1 Adaptation of the KoMus test

As a starting point, we used the short version of the 29-item KoMus test (see also 2.2.2). By using this already validated short version, and despite the reduction in test time to about 30 min, we were able to obtain a reliable test instrument that covered the four subdimensions of the test with a sufficient number of items. As described above, we initially wanted to use the MARKO test (Ehninger et al., 2021) to assess music-related argumentation competence; however, this test was not available for primary/lower secondary school students. As an alternative, we decided to include additional items from the KoMus test facets that dealt with music-related reflection. These items also served as a basis for the development of the MARKO test, and one of them was even included in the final MARKO test (see Ehninger et al., 2021). Therefore, the adapted Norwegian KoMus version was extended by four items and hence consists of 33 items. The adaptation procedure included the modernization of some of the questions (e.g., by changing the phrasing of "listening to a CD" to "listening to Spotify") and the alteration of some of the sound files to make them relevant in a Norwegian context (e.g., the use of a sound clip of children singing in Norwegian instead of German and the replacement of German folk music with Norwegian folk music).

In the following sections, we present a number of example items from the Norwegian KoMus version (translated into English for this paper) that illustrate what students see on their computer screens. When beginning the KoMus test, students are introduced to the functionality of the play button. This button is used to initiate the playback of the audio files. Furthermore, an alternative interface element featuring an icon resembling headphones is optionally provided. This particular tool enables students to opt for an auditory rendition of the textual content should they prefer such an approach (instead of reading the text).

Figure 2 provides an example item taken from dimension 1 (perception and musical memory). When the participants click the play button, they first hear a snare drum playing a rhythm twice. After that, they hear four different rhythms, one of them being the first rhythm they heard. The participants then decide which of the four rhythms is the same as the first rhythm and tick off the corresponding box.

Figure 3 provides an example item from dimension 2 (terminology and critical reflection). This item was adapted using a Norwegian child singing a pop song with a band. In this task, the participants are asked to give feedback on the girl singing, such as giving her advice on what she could improve. The participants press the play button to hear the audio file and then type their feedback in the corresponding field.

FIGURE 2
Screenshot of item example 1 from the KoMus test.

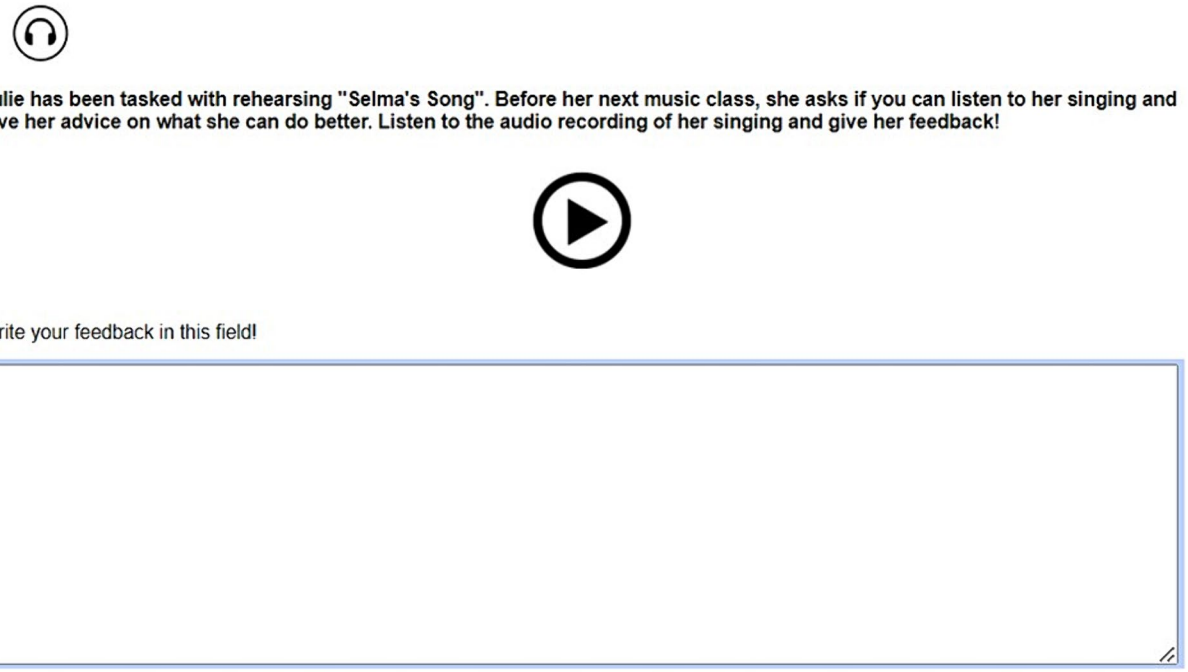

FIGURE 3
Screenshot of item example 2 from the KoMus test.

3.2 Adapted version of the KOPRA-M test

The original KOPRA-M test setup (Hasselhorn and Knigge, 2021) consists of a laptop, on which tasks are presented, a headset; an Android tablet with the Colored Music Grid (CMG) app (Hasselhorn and Grollmisch, 2014; see also Figures 1, 4) as a digital interface for playing melodies, accompaniments, and rhythms; and a local server for distributing all test items to the students' workplaces via a LAN. This setup has several disadvantages: the CMG app developed for Android has a clearly noticeable latency when playing music; putting together such a complex setup, with one laptop and one tablet per student and as a LAN with a server, is both expensive and time-consuming in terms of setup and take-down when testing; and, most importantly, it is very error-prone due to the multitude of devices and cable connections. Therefore, we decided to transfer the complete test setup to an iPad app as part of the adaptation procedure. Latency is known to be barely noticeable when performing music on iPads. In addition, the LAN became obsolete because the app could upload and download all necessary data via a web server.

The iPad version of the test—just like the original test—can be administered in a group setting in which each student is given an iPad and headphones. A researcher usually conducts the testing with

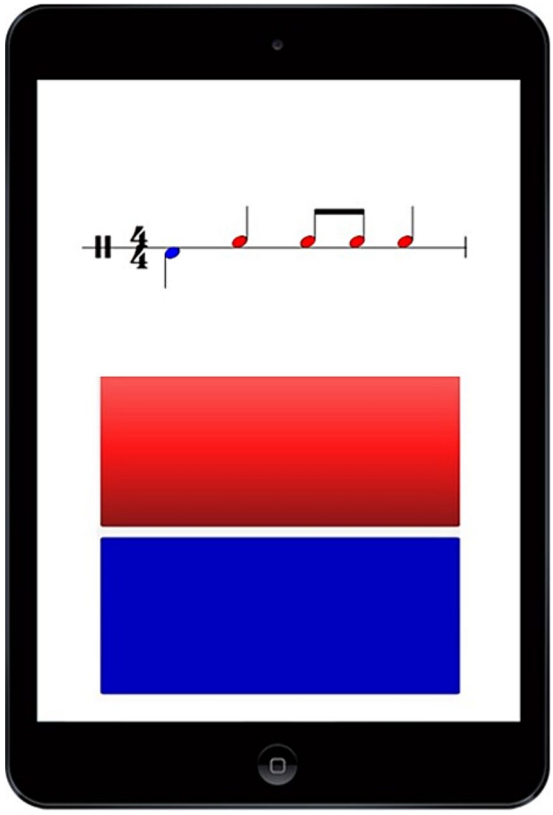

FIGURE 4
Item example from the KOPRA-M test (dimension: rhythm reproduction).

1–2 research assistants and provides the class with a set of iPads for this purpose. Recordings of the students playing and singing are stored directly in the app and uploaded to the server afterwards.

An example of such a task might be that students first hear a specific rhythm and subsequently play it on top of an accompaniment. In the singing dimension, there are tasks in which students sing their "answers"—for example, singing "Brother John" (originally "Frère Jacques") on top of an accompaniment—so that it can be determined whether the student is singing in tune and keeping rhythm and pace.

Before we recorded the participants' responses to the tasks, they were provided with a training session in which they tried out the CMG app as the first part of the implementation on the iPad. This introduction was programmed into the app as a short tutorial video. The following examples show what the rhythm and instrumental dimensions looked like on the iPad for the students. The text explains the audio instructions given to the participants.

In the first example (Figure 4), the students are asked to play the rhythms they heard using the red and blue fields, with each color corresponding to a different sound. Next, the students hear the rhythm represented by the notes twice. After that, they are to play this rhythm on top of an accompaniment, with the app recording the performance.

In example 2, the participants are given a melodic task (Figure 1). The students are asked to play the melody using the CMG. They hear the melody twice, with the option of playing along, after which the app records them playing the melody on top of an accompaniment.

In total, the Norwegian version of the KOPRA-M test consists of 21 items distributed across the three subdimensions as follows: song (four items), rhythm (10 items), and instrumental (seven items).

4 Data collection and participants

Data collection was carried out in three municipalities in southern and central Norway. The sample consisted of fifth graders (aged 9–10 years) from eight different public primary schools (KoMus: $N = 374$; 50% boys, 43% girls, and 7% other/did not answer; KOPRA-M: $N = 370$; 54% boys, 43% girls, and 3% other/did not answer). The participants' schools were invited to participate in the OutMus study.

The tests were conducted in two sessions on two different days. One session (maximum duration 60 min) was dedicated to conducting the KOPRA-M test. The other session (maximum duration 120 min) included both the KoMus test and other inventories relevant to the main study (OutMus; see Footnote 1). These tests were distributed via an online platform (SoSci Survey) and completed by the students via school-owned computers (Chromebooks). In the second test session, there was great variation regarding time spent, ranging from students completing the questionnaire in about 30 min to others not being able to complete the test within a timeframe of 120 min. As the KOPRA-M test is conducted in such a way that all students work on the same items at the same time, no variation regarding individual test times could occur during the first test session.

The tests were conducted during school hours in the participants' classrooms, with the students grouped according to their regular class assignments. Group sizes varied from 18 to 34 students. The students were seated individually at their own desks with a computer or iPad in front of them and headphones with a microphone attached.

Each test session started with a short oral introduction from the test leader. Additionally, a verbal introduction was played in the app itself to explain the items that were conducted there. In addition to a written explanation, all items in the questionnaire had a button to read the text aloud for those participants who wanted or needed to use this functionality.

Unfortunately, we had to deal with technical issues regarding the recording function of some iPads, which resulted in missing data from 202 participants on the singing tasks (KOPRA-M).

5 Data analysis

5.1 Coding process: rating of musical responses

We used the KOPRA-M rating scale to assess all the dimensions of the KOPRA-M test (Table 1). This rating scale is an adaptation of an assessment rubric for singing by Hornbach and Taggart (2005; for more details regarding the KOPRA-M adaptation, see Hasselhorn, 2015).

All data from the KOPRA-M test were coded by music experts. A workshop was conducted to ensure a common understanding of how to apply Hornbach and Taggart's (2005) rating scale. After the workshop, 210 recordings were independently scored by the expert raters, and interclass correlation coefficients (ICC) were used to assess

TABLE 1 Hornbach and Taggart's (2005) singing performance assessment rubric, adapted by Hasselhorn (2015, p. 78).

Rating	Description—Singing
5	The child sings the song nearly or completely accurately.
4	The child sings with some accuracy, beginning in the established key.
3	The child sings the song with some accuracy, starting in a different key than established, or modulates within the song.
2	The child sings/chants the melodic shape at a significantly different pitch.
1	The child sings/chants with a different melodic contour than the song.
0	Reasonable scoring is not possible (no signal or the child is not seriously involved with the task).

the interrater reliability. The analysis of a consistency two-way random effects model based on mean ratings (k = 3) yielded 95% confidence intervals ranging from 0.84 to 0.89 ($p < 0.001$). According to Koo and Li (2016), results between 0.75 and 0.90 indicate good reliability.

5.2 Statistical analysis procedures

The analyses were carried out using R (version 4.2.2; R Core Team, 2022) with the packages IRR (Gamer et al., 2022), eRm (Mair et al., 2021), psych (Revelle, 2023), TAM (Robitzsch et al., 2022), and IRT-scaling with ConQuest (version 5.29; Adams et al., 2022). Because the KoMus and KOPRA-M tests are both multidimensional, we used a generalized multidimensional partial credit Rasch model (Adams et al., 2023) to analyze the collected data. For computational reasons, we conducted analyses only for participants who provided values for at least three items per test subdimension. Missing values were not imputed.

We ensured that the standards related to classical test theory were met as proposed by Wu et al. (2016, p. 73–90). We also checked 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, weighted mean square (MNSQ) item fit indices were calculated considering conventional cut-off criteria (e.g., Bond and Fox, 2013; Ames and Penfield, 2015). In addition, classical item discrimination was determined as the point-biserial correlation of the item response category with the person ability (WLE) measured in the test.

The global fit of the model and the assumption of local stochastic independence were examined using Andersen's likelihood ratio test and information criteria (e.g., AIC and BIC). To ensure the fairness of the test, we conducted analyses of differential item functioning (DIF). When employing IRT, the probability of successfully answering an item is determined by an individual's ability. DIF is identified when other factors (e.g., the person's gender) influence the probability of individuals with the same abilities solving an item.

To establish the credibility of the KoMus and KOPRA-M tests, we examined their validity through both convergent and discriminant validity analyses. Convergent validity, as described by Gregory (2015, p. 130), is demonstrated when two tests measuring the same construct exhibit a high correlation, while discriminant validity is indicated when tests measuring different constructs show negligible correlations. Cohen's (1988) criteria for interpreting correlation coefficients suggest that a coefficient of 0.10 represents a small correlation, 0.30 represents a medium correlation, and 0.50 or higher represents a large correlation.

To assess the validity of the adapted competency tests, we calculated Pearson's correlation coefficients (r) to analyze the linear associations between the dimensions of the KOPRA-M and KoMus tests. In addition, we examined the correlations between these competency tests and nonverbal intelligence, which were measured using a short digital version of Raven's 2 test[2] (Raven, 2018) as an unrelated construct. Initially, a Kolmogorov–Smirnov test was employed to examine the normality of the data. This step was crucial to fulfilling the prerequisites for utilizing Pearson's r.

6 Results

6.1 Psychometric evaluation of the KoMus test

Based on the procedure described above, five items were eliminated from the adapted KoMus test, having yielded unacceptable values on several indices (item fit and discrimination). After elimination, 28 items were used for further analysis. Table 2 shows the relevant statistical information regarding the remaining item pool.

We conducted DIF analyses with the group variable gender. We followed the categorization proposed by the Educational Testing Service, assuming that an effect size of 0.64 logits or greater indicates moderate to large DIF (Trendtel et al., 2016, p. 131), and we found no DIF for any of the items.

The global fit of the model and the assumption of local stochastic independence were examined using Andersen's likelihood ratio test. This showed a nonsignificant result when the sample was split into two subsamples using a random split criterion (even vs. uneven case number, $p = 0.92$) and gender as a split criterion (male vs. not male, $p = 0.86$). Therefore, we assumed that the IRT model used would fit our data.

To check whether the dimensional structure of the original test could also be applied to our data, we tested different models against each other using different information criteria (AIC, AICc, and BIC; the smaller the information criteria, the more efficiently the corresponding model explains the observed data; e.g., Bozdogan, 1987). The starting model for the comparison was unidimensional (Model A), with the assumption of a single latent variable. Model B corresponded to the final four-dimensional KoMus model. Model C corresponded to the original theoretical KoMus model, in which the "critical evaluation of music and its performance" was defined as a separate dimension, in addition to the four dimensions of Model B (Jordan et al., 2012, pp. 503–504). Furthermore, models D and E were used to investigate whether the KoMus dimension of "perception and musical memory"

2 The results of the Raven's 2 test are published with permission from Pearson Sweden AB. Pearson was not involved in the study and is not responsible for either the quality of the results or the overall outcome of the conducted research. Copyright © 2018 NCS Pearson, Inc. Norwegian translation copyright © 2020 NCS Pearson, Inc. Adapted and reproduced by Pearson Sweden AB under license from Pearson Inc. All rights reserved. Pearson and Raven's are trademarks, in the US and/or other countries, of Pearson Education, Inc., or its affiliates.

TABLE 2 Summary of the most important psychometric characteristics of the items of the KoMus test.

Dimension[a]	No. of items	Item difficulty (classical)		Item difficulty (IRT)	Item fit (MNSQ)		Item discrimination		Reliability (EAP/PV)
		Min/Max	*M* (SD)	Min/Max	Min/Max	*M* (SD)	Min/Max	*M* (SD)	
1. (perception and musical memory)	7	17.82/71.28	41.85 (21.07)	−1.20/0.92	0.93/1.10	0.99 (0.06)	0.22/0.43	0.33 (0.08)	0.674
2. (terminology)	6	3.2/69.48	25.09 (24.51)	−2.71/2.32	0.97/1.06	1.00 (0.04)	0.26/0.51	0.39 (0.10)	0.785
3. (notation)	6	13.41/59.94	35.40 (19.72)	−1.35/1.38	0.93/1.09	1.01 (0.06)	0.24/0.51	0.35	0.678
4. (contextual knowledge)	3	5.03/64.35	43.67 (33.49)	−1.37/2.23	0.94/1.09	1.01 (0.05)	0.35/0.41	0.39 (0.03)	0.605
5. (critical reflection)	6	5.5/17.65	9.17 (4.72)	0.07/3.18	0.95/1.08	1.03 (0.05)	0.28/0.52	0.41 (0.11)	0.730

[a]We have used the dimensional structure here because it results from the dimensionality check (see Table 3).

TABLE 3 Information criteria for the estimated KoMus models.

Model	*N*	Parameter	Deviance	AIC	AICc	BIC
A – unidimensional	376	43	12822.0	12908.0	12919.4	13077.0
B – 4 dimensions	376	52	12723.7	12827.7	12844.7	13032.0
C – 5 dimensions	376	57	12639.3	12753.3	12774.1	12977.2
D – 4 dimensions + main dimension	376	52	12779.9	12883.9	12901.0	13088.2
E – 5 dimensions + main dimension	376	57	12732.8	12846.8	12867.6	13070.8

TABLE 4 Summary of the most important psychometric characteristics of the items of the KOPRA-M test (item fit and item difficulty estimated at the item category level).

Dimension	No. of items	Item difficulty (classical)		Item difficulty (IRT)	Item fit (MNSQ)		Item discrimination		Reliability (EAP/PV)
		Min/Max	*M* (SD)	Min/Max	Min/Max	*M* (SD)	Min/Max	*M* (SD)	
1. (instrumental play)	7	7.58/49.00	23.97 (13.66)	−3.34/3.05	0.81/1.12	0.98 (0.06)	0.39/0.53	0.45 (0.05)	0.765
2. (singing)	4	10.06/30.18	18.89 (6.41)	−2.38/2.80	0.82/1.23	0.97 (0.10)	0.59/0.62	0.60 (0.02)	0.505
3. (rhythm production)	10	4.97/69.23	24.67 (19.19)	−3.96/2.67	0.70/1.30	0.97 (0.96)	0.53/0.67	0.60 (0.05)	0.881

could be regarded as a main dimension (general factor) to which the other dimensions would be subordinate (for this purpose, a subdimension model was estimated using ConQuest; Brandt, 2008).

As shown in Table 3, Model C was most likely the best model.

6.2 Psychometric evaluation of the KOPRA-M test

No items were eliminated from the adapted KOPRA-M test. Table 4 shows the relevant statistical information regarding the item pool.

The same procedure described for the KoMus test was followed for the detection of DIF; similarly, no DIF was found for any of the items.

Andersen's likelihood ratio test yielded a nonsignificant result when the sample was split into two subsamples using a random split criterion (even vs. uneven case number, $p = 0.57$) and gender as a split criterion (male vs. not male, $p = 0.09$). Therefore, we assumed that the IRT model used would fit our data.

We once again followed the procedure of the test's authors (Hasselhorn, 2015) and used information criteria to compare a unidimensional model (A) against a three-dimensional model (B) (see the more detailed description in the Psychometric Evaluation of the KoMus Test section). As shown in Table 5, Model B was most likely the best model, meaning that the dimensionality of the Norwegian variant was identical to that of the original KOPRA-M test.

6.3 Validity of the adapted competency tests

We hypothesized that the dimensions of the KoMus test, which are perception based, would exhibit large correlations with each other.

TABLE 5 Information criteria for the estimated KOPRA-M models.

Model	*N*	Parameter	Deviance	AIC	AICc	BIC
A – unidimensional	371	95	16089.6	16279.6	16345.9	16651.6
B – 3 dimensions	371	100	15368.5	15568.5	15643.3	15960.1

TABLE 6 Correlations between the different dimensions of the KOPRA-M and KoMus tests.

	KOPRA D1	KOPRA D2	KOPRA D3	KoMus D1	KoMus D2	KoMus D3	KoMus D4	KoMus D5
KOPRA D1	1	.	.	.	.	.	.	.
KOPRA D2	0.41	1	.	.	.	.	.	.
KOPRA D3	0.51	0.64	1	.	.	.	.	.
KoMus D1	0.42	0.35	0.48	1	.	.	.	.
KoMus D2	0.44	0.37	0.50	0.98	1	.	.	.
KoMus D3	0.43	0.37	0.49	0.94	0.96	1	.	.
KoMus D4	0.39	0.32	0.44	0.85	0.90	0.77	1	.
KoMus D5	0.42	0.32	0.44	0.82	0.89	0.78	0.93	1

KoMus dimensions: 1 = perception & musical memory, 2 = terminology, 3 = notation, 4 = contextual knowledge, and 5 = critical reflection; KOPRA-M dimensions: 1 = instrumental play, 2 = singing, and 3 = rhythm production. All correlations have a significance level at $p < 0.001$.

TABLE 7 Correlations between competency dimensions and IQ.

	KOPRA D1	KOPRA D2	KOPRA D3	KoMus D1	KoMus D2	KoMus D3	KoMus D4	KoMus D5
IQ	0.20***	0.05	0.18**	0.24***	0.26***	0.25***	0.28***	0.25***

** $p < 0.01$, *** $p < 0.001$.

Similarly, we expected medium to large correlations between the dimensions of the KOPRA-M test and between the KoMus and KOPRA-M tests, given their theoretical similarities.

Prior research has shown significant correlations between IQ and musical abilities (Schellenberg and Lima, 2024, p. 21), leading us to anticipate small correlations between these constructs. Based on this framework, we expected correlations in the discriminant validity analyses to be less than 0.30 and correlations in the convergent validity analyses to be greater than 0.30.

In Table 6, notable significant correlations can be seen among the five dimensions of the KoMus test, while medium to large significant correlations exist among the three dimensions of the KOPRA-M test and between the dimensions of the KoMus and KOPRA-M tests. This demonstrates substantial convergent validity according to the theoretical models behind the two tests. While all the dimensions of the KoMus test are perception based and strongly linked to each other, the dimensions of the KOPRA-M test are not. However, it remains clear that both tests measure related musical competencies.

Table 7 shows that only small correlations were evident between IQ scores and all dimensions in both competency assessments, demonstrating discriminant validity. Notably, the weakest correlation was identified between IQ and the singing dimension of the KOPRA-M test; this was also the only nonsignificant relationship.

The outcomes of the normality test were nonsignificant, allowing us to assume that both the competency test variables and the IQ test scores adhered to a normal distribution and would therefore be suitable for the validity analyses.

6.4 Tests and models "in practice"

Administering these tests to fifth-grade students provided us with valuable insights into the assessments' practical benefits. One particularly notable aspect, as highlighted in Section 6.4.1, is the capacity to explore a student's competence structure. This entails a comprehensive evaluation of a student's strengths and the areas in which they may require improvement. Just as a high level of musical competence does not guarantee proficiency in all musical aspects, our competency test emphasizes the fact that expertise is multifaceted. This aligns with the Norwegian music curriculum, which provides room for the exploration of several areas and levels of musical competence as described both in the subject's relevance and central values and core areas (Kunnskapsdepartementet, 2019).

The competency test uniquely allowed us to conduct a nuanced analysis of the students' competencies. This level of detail opens up a range of opportunities for tailored didactic applications, enabling the development and support of specific competencies through targeted teaching methods in line with the curriculum that clearly sets out to foster the development and growth of musical competencies. Such knowledge could have implications for how teachers assess musical competencies over time in the classroom.

Another valuable aspect, as detailed in Section 6.4.2, is the ability to provide precise descriptions of competency content. This involves offering a detailed account of the specific tasks that students can effectively address and resolve.

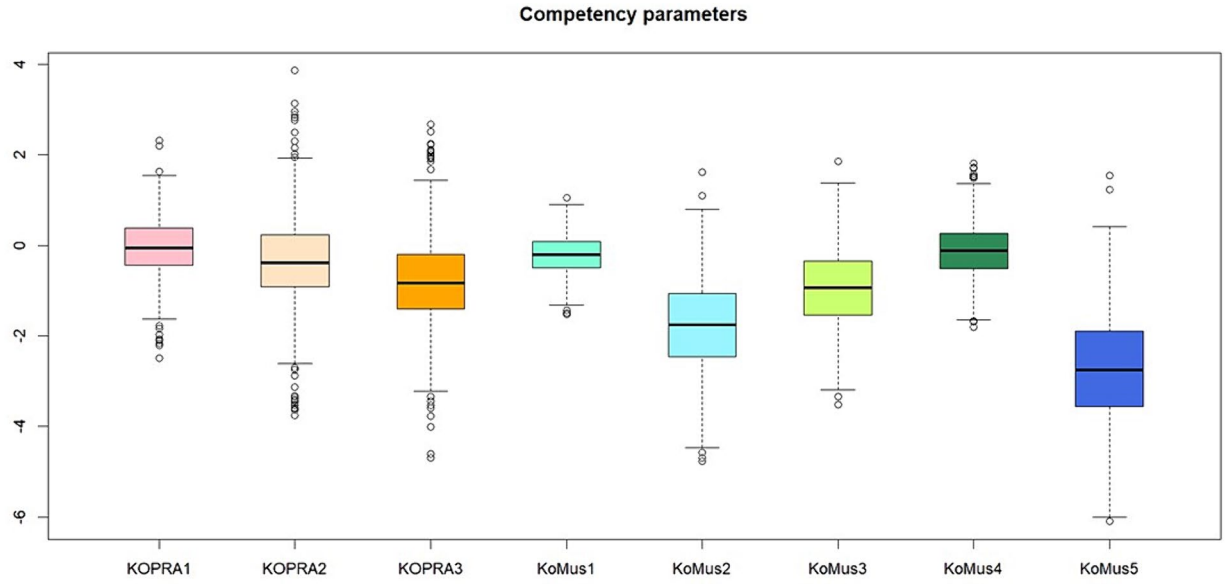

FIGURE 5
Boxplots showing the distribution of the KoMus and KOPRA-M dimensions.

6.4.1 Students' competency structures

Figure 5 illustrates that the most challenging dimension was dimension 5 in the KoMus test, focusing on critical reflection, with a mean students' competency estimate ($\overline{\theta}$) of −2.75 logits and a standard deviation (SD) of 1.26 logits. Following closely was dimension 2 in the KoMus test, pertaining to terminology ($\overline{\theta}$ = −1.76; SD = 1.08). Notably, dimension 5 in the KoMus test exhibited the highest variance, indicating substantial variability in the participants' competency levels within this dimension. Similarly, the singing dimension in the KOPRA-M test indicated a wide range of competency ($\overline{\theta}$ =−0.33; SD = 1.21).

Conversely, the dimensions in which participants scored the highest were dimension 1 in the KOPRA-M test, emphasizing instrumental play ($\overline{\theta}$ =−0.068; *SD* = 0.69), and dimension 4 in the KoMus test, focusing on contextual knowledge ($\overline{\theta}$ =−0.13; *SD* = 0.63).

6.4.2 Students' competency levels

In the original publication on the KOPRA-M assessment (Hasselhorn, 2015), several competency levels were established and described. To provide an example of the potential of an IRT-based test instrument for music education in Norway, we conducted a similar analysis of the rhythm production dimension of the KOPRA-M test based on our empirical material. We used a Wright map (Figure 6) to visualize the relationship between a person's competence and the difficulty of a particular item. The distribution of the item categories was ordered by 65% solution probability. A histogram depicting the distribution of proficiency scores (i.e., students' competences) is shown on the left side of the diagram, while on the right side, the difficulty levels of the item categories are displayed (Figure 6).

The original competency levels of the KOPRA-M test were described for the dimension of rhythm production by Hasselhorn (2015, p. 140–153). Adapted to our data, the competency-level descriptions can be formulated as follows:

Level 1. Scale points from −0.5 to 0.75. Students at this level have the following competency:

- They can play rhythm patterns (e.g., pattern 41, Figure 7) at their own pace and with several mistakes when these patterns are demonstrated to them at a moderate tempo (*ca.* 90 bpm) and when corresponding sheet music examples are provided.

Level 2. Scale points from 0.75 to 2.0. Students at this level can achieve the following:

- Play rhythm patterns without syncopation or ties across the bar line (e.g., pattern 411, Figure 7) almost flawlessly in the given tempo if these patterns have been played to them and the corresponding sheet music has also been presented.
- Play rhythm patterns without syncopation or over-tying with some errors at the given tempo when these patterns have been played to them but no scores have been presented.

Level 3. Scale points greater than 2.0. Students at this level can achieve the following:

- Play rhythm patterns (e.g., pattern 42, Figure 7) almost or completely flawlessly at the given tempo when these patterns have been played to them.
- Play rhythm patterns almost or completely flawlessly at the given tempo when presented with corresponding sheet music examples.
- Play rhythm patterns with syncopation or over-tying (e.g., pattern 45, Figure 7) with some mistakes.

In our sample, this resulted in the following distribution: 40% of the participants performed at competency level 1, 23% at level 2, and 6% at the highest level (level 3); the remaining 31% performed below level 1. This indicates that while the majority of the fifth graders participating in this study had a competency level that allowed them to perform rhythms at their own pace, with some mistakes after hearing them and seeing the corresponding notation on sheet music,

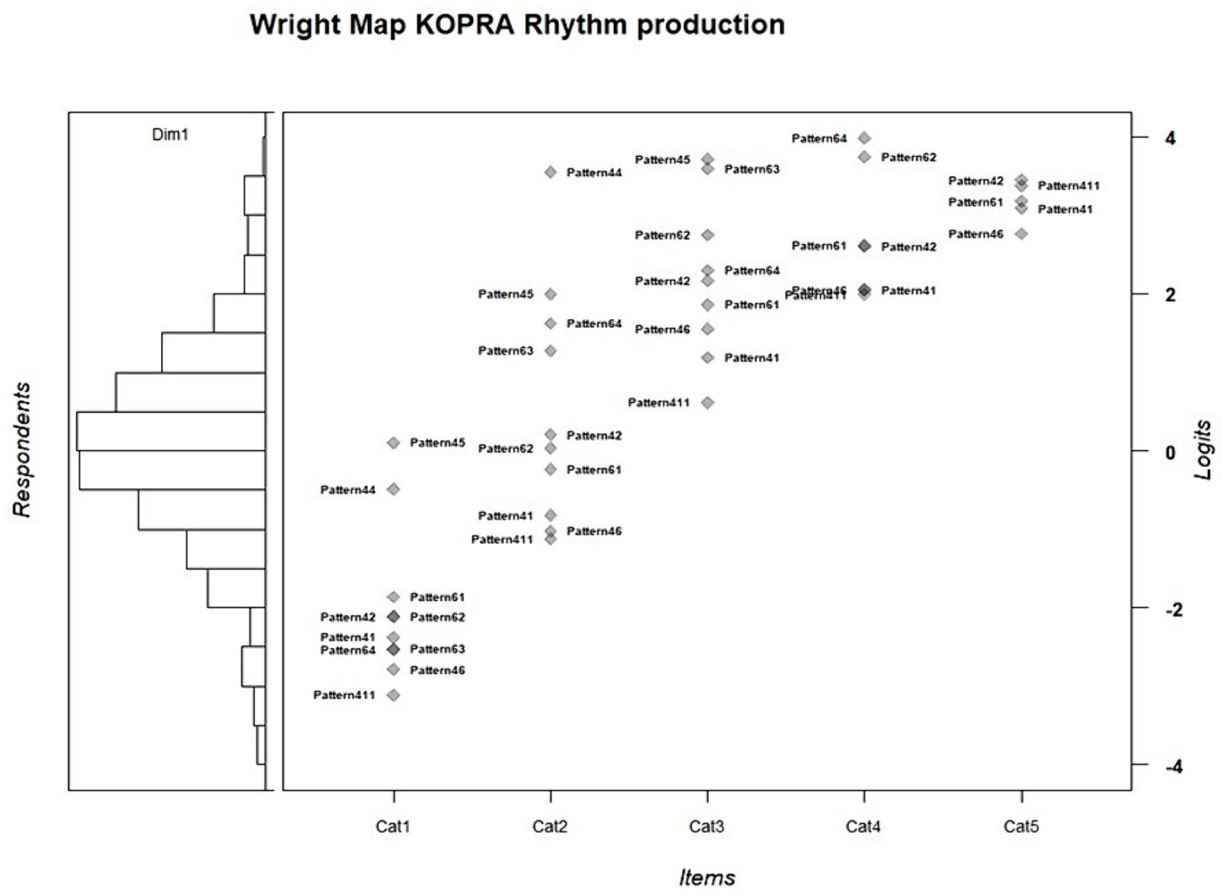

FIGURE 6
Wright map for all items belonging to the KOPRA-M test's rhythm production dimension.

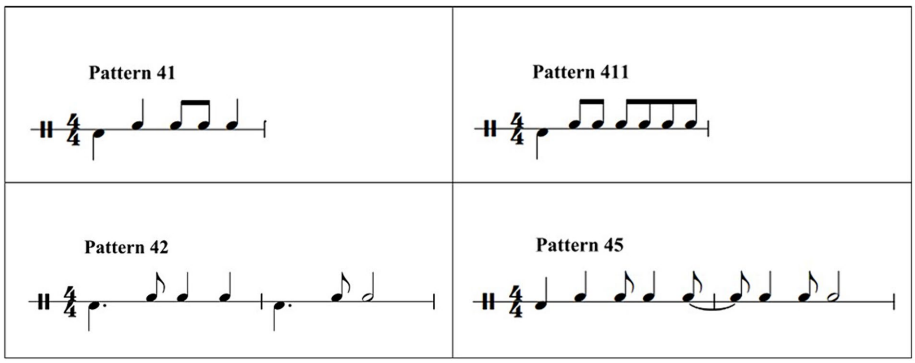

FIGURE 7
Rhythm patterns from the KOPRA-M test.

almost one-third of them did not yet have this competency. At a higher level, students could also maintain a steady tempo, even without visual cues. A smaller number of students performed at the highest level and could handle more complex rhythms while maintaining a consistent tempo. Even though some individuals were able to reach competency level 3 in our sample, none of the participants were able to perform, for example, pattern 45 at the highest scoring level.

7 Discussion

7.1 Results summary

The main aims of this study were to select, adapt, and validate assessments that are well-suited for evaluating the learning outcomes (competencies) of music lessons in Norwegian primary schools. Assessing music competence in alignment with the curriculum

TABLE 8 EAP/PV reliability coefficients from three studies using the KOPRA-M assessment.

	D1 (instrumental play)	D2 (singing)	D3 (rhythm production)
Hasselhorn (2015)	0.96	0.91	0.92
Lill et al. (2019)	0.78	0.86	0.81
Current study	0.77	0.51	0.88

remains a developing area and is particularly uncharted in Norway, where empirical (and especially quantitative) investigations into such competencies are absent. To achieve these objectives, the KoMus and KOPRA-M assessments were selected based on their consideration of factors such as age appropriateness, accessibility, technological state of the art, curricular relevance, and psychometric qualities. We decided to use short versions of these assessments that allow for shorter test times and that can be used in studies in which other variables/constructs are also of interest (which is almost always the case).

After collecting data from a sample of Norwegian fifth graders, we conducted comprehensive analyses of the psychometric characteristics of both tests. During this process, we identified and removed five items from the KoMus test due to poor item fit and item discrimination. Consequently, the adapted version of this test comprises 28 items. For the KOPRA-M assessment, all 21 items could be used for the Norwegian version.

7.2 Discussion of psychometric characteristics

7.2.1 Fit and dimensionality

Using a generalized multidimensional partial credit Rasch model, we confirmed that the test instruments demonstrated an appropriate overall fit and confirmed the assumption of local stochastic independence. Nonetheless, upon detailed examination, we found areas that are worth discussing further.

Our empirical findings revealed that the selected items from the KoMus assessment represent a five-dimensional competency construct that differs from the original four-dimensional structure (Jordan et al., 2012) and aligns with our decision to incorporate additional items from the KoMus test facets pertaining to music-related reflection. While in the original study (Jordan et al., 2012), Model C was the second best and Model B was (by a narrow margin) the best model, the order was reversed in our case. Jordan et al. (2012) originally expected this result (i.e., Model C being the best-fitting model) based on their theoretical assumptions. Thus, while our study did not reveal a fundamentally new structure, it suggests that the number of items per dimension has an influence. While in Jordan et al. (2012), the "critical evaluation" dimension was represented using significantly fewer items compared to the other dimensions, this was not the case in our study (see Table 2). Against this background, we propose the use of the KoMus test version adapted for Norway as a five-dimensional test. This result has several consequences. First, it limits the direct comparability of the results of Norwegian and German studies. However, this seems negligible to us, because no international comparative large-scale assessments in music are planned in the foreseeable future. Should this be the case, however, the four-dimensional use of the Norwegian KoMus version is possible at any time—both at the item and dimensional levels—as all relevant psychometric characteristics of the four-dimensional version are also within an acceptable to good range (and very similar to those of the five-dimensional version).[3] Second, and this is a positive consequence, the use of a five-dimensional version results in the possibility of measuring and reporting students' competencies in an even more differentiated way than before.

7.2.2 Reliability

In the competency domain of perception-based contextual musical knowledge (KoMus dimension 4), lower reliability (0.605) than in the original test version (Jordan et al., 2012) was observed. This was also the case with the second dimension of the KOPRA-M test (0.505), the subdomain of singing (Hasselhorn, 2015). Several factors might explain this phenomenon. While the KoMus and KOPRA-M assessments exhibit exceptional precision as measurement instruments, the Norwegian curriculum sets forth very broad competency objectives. It is therefore plausible that some students might lack the required knowledge to successfully engage with these items, leading them to guess more in some test dimensions than in others, which in turn leads to lower reliability. This also indicates that these assessments may be valid in a research context but may not be suitable as an evaluation tool within specific school contexts (i.e., at the class and individual levels). We generally observed lower reliability in the Norwegian short versions than in the original German tests. This was expected, considering that a smaller number of items usually decreases reliability. For instance, the singing dimension from the KOPRA-M test consists of only four items in the adapted version. For comparison, the reliability coefficients from two studies utilizing the KOPRA-M test (Hasselhorn, 2015; Lill et al., 2019) and the current study are shown in Table 8. Furthermore, the analysis of the KoMus test yielded a structure consisting of five dimensions, decreasing the number of items belonging to each dimension and lowering the reliability of single test dimensions. Still, we found the reliability of both tests to be within an acceptable range.

7.3 Practical advice and possible further developments

During the data collection process, we observed significant variations in the time the participants allocated to the KoMus assessment. While some participants invested substantial effort in crafting detailed, extensive responses, others occasionally became impatient during the sessions and left the tasks incomplete. Notably, our observations suggest that individuals with lower competency levels may encounter particular challenges due to the test's length.

3 The psychometric characteristics of the four-dimensional version are available upon request from the authors.

We would like to suggest two ways to deal with this issue. First, in terms of practical advice, future test users could use an even shorter version of the KoMus test. Based on our analysis, it would be possible to reduce the test to 10 items while still ensuring the psychometric qualities of the instrument. However, a disadvantage would be that, with a limited number of items allocated to each dimension, the subdimensions could no longer be included in the differential analyses. Second, another possibility is to conduct the test adaptively. The item parameters generated by our study and the adaptive test platforms available today should allow such a procedure in principle. However, all open items would then have to be omitted (at least until an AI model has been trained for automated rating). Since the items in question belong primarily to the dimension of critical reflection (D5), this dimension could no longer be measured—or at least not with a high degree of precision.

7.4 Limitations

Although our study yielded several positive outcomes, it is important to recognize and address its limitations. One limitation pertains to the time and effort required to rate responses in the KOPRA-M assessment. However, it is very likely that in the future, advancements in AI or machine learning will streamline this process, significantly reducing the hours and labor needed to generate ratings and provide instant results.

Additionally, this study raises questions about other factors that might influence the development of music-related competencies and the fairness of the test. Our validity analysis indicated small, yet significant, relationships between IQ and several dimensions of the assessments. Such relationships warrant further exploration to gain a deeper understanding of how, for example, cognitive abilities and reading comprehension can affect participants' performance on such tests. Additionally, factors such as participants' socioeconomic status, family interest in music and musical background could influence the fairness of the tests. These aspects were not explored in the current sample, which leaves questions about the representativeness of the sample open for further discussion.

It is worth noting that while the assessments encompass several fundamental domains of the Norwegian curriculum, they do not address the domain of music creation. Therefore, the adapted assessments cannot be considered a comprehensive test battery that fully evaluates music competence in Norwegian primary/lower secondary schools. An important future endeavor should therefore involve developing or adapting an assessment specifically designed to assess competencies in music creation.

7.5 Conclusion

Our findings serve as an important first step in exploring musical competencies in Norwegian primary schools. The Norwegian adaptions of the KoMus and KOPRA-M assessments exhibit robust psychometric properties. They offer a valuable tool for delving deeper into musical competency structures, as shown through an illustrative example in which one of the dimensions of the IRT-based test was used to model competency levels. In the future, such approaches can allow for a better understanding of learning processes and the support of competency development in the areas of musical performance, perception, and contextualization. This also opens up opportunities for conducting longitudinal studies and exploring factors relevant to the development of musical competencies, as we are currently doing in our forthcoming article (Nørstebø and Knigge, 2024). Ultimately, these efforts will facilitate evidence-based professional and educational development.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by SIKT Norwegian Agency for Shared Services in Education and Research. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

KN: Writing – original draft, Writing – review & editing. JK: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This paper was produced as part of the OutMus project funded by the Research Council of Norway (RCN/NFR; grant no. 320141).

Acknowledgments

We would like to thank all the collaborating partners who made this study possible. These include all the dedicated teachers, pupils, primary schools, and schools of music and performing arts that took part in the OutMus project. We also thank Kristine Antun, Jean Sebastian Aubert, Anna-Kristine Erichsen, Therese Hagir, Ingrid Lauten, and Abelone Torp for their assistance in data collection and data handling. Anders Rønningen and Norsk kulturskoleråd supported the study, facilitating collaboration with all the schools. Johannes Hasselhorn, Andreas C. Lehmann, and Florian Lill provided us with the original German KOPRA-M version, and Jens Knigge and Anne-Katrin Jordan provided the KoMus test material. Finally, we thank Ingmar Baetge for programming the KOPRA-M iPad app.

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

References

Adams, R. J., Cloney, D., Wu, M. L., and Osses, A. V. S. (2023). ACER ConQuest manual. Australian council for. *Educ. Res.*

Adams, R. J., Cloney, D., Wu, M., Osses, A., Schwantner, V., and Vista, A. (2022). ACER ConQuest manual. Australian Council for Educational Research. Camberwell. Available at: https://research.acer.edu.au/measurement/5

Allen, N. L., Jenkins, F., and Schoeps, T. L. (2004). The NAEP 1997 arts technical analysis report (Report No. ETS-NAEP 04-T01). Educational Testing Service. Available at: http://www.ets.org/Media/Research/pdf/ETS-NAEP-04-T01.pdf

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, 39–48. doi: 10.1111/emip.12067

Bond, T. G., and Fox, C. M. (2013). Applying the Rasch model: fundamental measurement in the human sciences. New York: Psychology Press.

Boyle, J. D., and Radocy, R. E. (1987). Measurement and evaluation of musical experiences. New York: Schirmer Books.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): the general theory and its analytical extensions. *Psychometrika* 52, 345–370. doi: 10.1007/BF02294361

Brandt, S. (2008). Estimation of a Rasch model including subdimensions. In DavierM. von and D. Hastedt (Eds.), Issues and methodologies in large scale assessments (Vol. 1, pp. 51–70). Hamburg: IEA-ETS Research Institute.

Brophy, T. (2019). The Oxford handbook of assessment policy and practice in music education, vol. *1*. New York: Oxford University Press.

Buerger, S., Kroehne, U., and Goldhammer, F. (2016). The transition to computer-based testing in large-scale assessments: investigating (partial) measurement invariance between models. *Psychol. Test Assess. Model.* 4, 597–616.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Colwell, R. (1999). The 1997 assessment in music: red flags in the sunset. *Arts Educ. Policy Rev.* 100, 33–38. doi: 10.1080/10632919909605996

Correia, A. I., Vincenzi, M., Vanzella, P., Pinheiro, A. P., Lima, C. F., and Schellenberg, E. G. (2022). Can musical ability be tested online? *Behav. Res. Methods* 54, 955–969. doi: 10.3758/s13428-021-01641-2

COTAN. (2019). COTAN review system for evaluating test quality. Available at: https://www.cotan.nl/review-system

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

Gamer, M., Lemon, J., and Singh, I. (2022). IRR: various coefficients of interrater reliability and agreement (version 0.84.1). The comprehensive R archive Network. Available at: https://CRAN.R-project.org/package=irr

Gembris, H. (1998). "Fundamentals of musical talent and development" in Forum Musikpädagogik, vol. *20* (Augsburg: Wißner).

Gregory, R. J. (2015). Psychological testing: History, principles, and applications. *7th* Edn. Boston: Pearson.

Harnischmacher, C., and Knigge, J. (2017). Motivation, music-making practice and interest in music in the family as predictors of the competence to perceive and contextualize music and of the experience of competence in music lessons. Contributions to Empirical Music Education, 8, 1–21. Available at: https://www.b-em.info/index.php/ojs/article/view/136

Hartig, J., and Klieme, E. (2006). "Competence and competence diagnostics" in Performance and performance diagnostics. ed. K. Schweizer (Heidelberg: Springer Medizin), 127–143.

Hasselhorn, J. (2015). *Messbarkeit musikpraktischer Kompetenzen von Schülerinnen und Schülern*: *Entwicklung und empirische Validierung eines Kompetenzmodells*. New York: Waxmann Verlag.

Hasselhorn, J., and Grollmisch, S. (2014). The colored music grid (CMG) app. A new input interface for capturing instrument-independent instrumental performance. In W. Auhagen, C. Bullerjahn and GeorgiR. von (Eds.), Music psychology. Yearbook of the German Society for Music Psychology. Volume 24: Open-earedness. A postulate in focus. Göttingen: Hogrefe.

Hasselhorn, J., and Knigge, J. (2021). "Technology-based competency assessment in music education: the KOPRA-M and KoMus tests" in *Testing and feedback in music education – Symposium Hannover 2017)*. eds. A. Lehmann-Wermser and A. Breiter, (ifmpf). Hannover.

Hornbach, C. M., and Taggart, C. C. (2005). The relationship between developmental tonal aptitude and singing achievement among kindergarten, first-, second-, and third-grade students. *J. Res. Music. Educ.* 53, 322–331. doi: 10.1177/002242940505300404

Jordan, A.-K., Knigge, J., Lehmann, A. C., Niessen, A., and Lehmann-Wermser, A. (2012). Entwicklung und Validierung eines Kompe-tenzmodells im Fach Musik–Wahrnehmen und Kontextualisieren von Musik. *Zeitschrift Für Pädagogik* 4, 500–521. Available at: http://nbn-resolving.de/urn:nbn:de:0111-pedocs-103923

Juntunen, M. L. (2017). National assessment meets teacher autonomy: national assessment of learning outcomes in music in Finnish basic education. *Music. Educ. Res.* 19, 1–16. doi: 10.1080/14613808.2015.1077799

Jurecka, A. (2008). "Introduction to the computer-based assessment of competencies" in Assessment of competencies in educational contexts. eds. D. L. J. Hartig and E. Klieme (Göttingen: Hogrefe & Huber), 193–214.

Klieme, E., and Leutner, D. (2006). Competency models for the assessment of individual learning outcomes and for the assessment of educational processes. Description of a newly established DFG priority Programme. *J. Educ.* 52, 876–903.

Koo, T. K., and Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* 15, 155–163. doi: 10.1016/j.jcm.2016.02.012

Kunnskapsdepartementet. (2019). Læreplan i musikk (MUS01-02). Available at: https://www.udir.no/lk20/mus01-02

Lill, F., Hasselhorn, J., and Lehmann, A. C. (2019). "The relationship between musical self-concept and practical music competencies in secondary schools" in Praxen und Diskurse aus Sicht musikpädagogischer Forschung (pp. 171–187). eds. V. Weidner and C. Rolle (New York: Waxmann). doi: 10.25656/01:20711

Mair, P., Hatzinger, R., and Maier, M. J. (2021). eRm: Extended Rasch modeling (Version 1.0-2). The Comprehensive R Archive Network. Available at: https://cran.r-project.org/package=eRm

National Assessment Governing Board. (2016). 2016 arts education assessment framework. Available at: https://www.nagb.gov/content/dam/nagb/en/documents/publications/frameworks/arts/2016-arts-framework.pdfwork

Nørstebø, K., and Knigge, J. (2024). A quasi-experimental study exploring the development of musical competence in children and factors that influence this development: Department for Arts and Culture. Levanger: Nord University.

R Core Team (2022). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing.

Raven, J. C. (2018). Raven's 2 progressive matrices – Clinical edition. Minneapolis: NCS Pearson, Inc.

Revelle, W. (2023). Psych: procedures for psychological, psychometric, and personality research. (version 2.3.6). The Comprehensive R Archive Network. Available at: https://CRAN.R-project.org/package=psych

Robitzsch, A., Kiefer, T., and Wu, M. (2022). TAM: test analysis modules (version 4.1-4). The comprehensive R archive Network. Available at: https://CRAN.R-project.org/package=TAM

Schellenberg, E. G., and Lima, C. F. (2024). Music training and nonmusical abilities. *Annu. Rev. Psychol.* 75, 87–128. doi: 10.1146/annurev-psych-032323-051354

Strauss, H., Reiche, S., Dick, M., and Zentner, M. (2024). Online assessment of musical ability in 10 minutes: development and validation of the Micro-PROMS. *Behav. Res. Methods* 56, 1968–1983. doi: 10.3758/s13428-023-02130-4

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

Weinert, F. E. (1999). Concepts of competence: DeSeCo expert report . Neuchatel: DeSeCo.

Weinert, F. E. (2001). "Concept of competence: a conceptual clarification" in Defining and selecting key competencies. eds. D. S. Rychen and L. H. Salganik (Seattle: Hogrefe), 45–65.

Wu, M., Tam, H. P., and Jen, T.-H. (2016). Educational measurement for applied researchers. Singapore: Springer.