Thinking in Music: An Objective Measure of Notation-Evoked Sound Imagery in Musicians

Anna Wolf[1,2], Reinhard Kopiez[1], Friedrich Platz[3]

[1] Hanover Music Lab, Hanover University of Music, Drama and Media, Hanover, Germany

[2] Institute for Systematic Musicology, Universität Hamburg, Germany

[3] University of Music and Performing Arts, Stuttgart, Germany

Update: 7 August 2018

Wordcount: XYZ

Author contact

Anna Wolf, Institute for Systematic Musicology, Neue Rabenstr. 13, 20354 Hamburg, Germany

Email: anna.wolf@uni-hamburg.de

Reinhard Kopiez: reinhard.kopiez@hmtm-hannover.de

Friedrich Platz: friedrich.platz@mh-stuttgart.de

Keywords: Auditory imagery, inner hearing, audiation, test construction, reading skill

Abstract

Being able to imagine the sound of music from notation as a result of so-called notational audiation, without the physical presence of sound, is an indispensable skill for professional musicians. However, up until now there has been no assessment for the evaluation of the skill responsible for the reading-imagery task. The development of an assessment for the skill of using notation-evoked sound imagery (henceforth called NESI) described in this study was based on the embedded melody paradigm. A large number of tasks were developed and evaluated for task difficulty and internal consistency. Participants first had to read a figural variation from notation and imagine the sound without singing or humming. After the notation disappeared from the screen, the sound of a theme was played which either matched the original melody harmonically or contained small but significant deviations from the original variation's harmonic progression (the so-called "lure" variation). Participants had to decide whether the variation and theme matched in terms of their harmonic structure. Starting from a large number of various items, we analyzed the item characteristics in a series of pilot studies for the selection of a reduced and validated number of item triads. Internal validity and unidimensionality were tested by use of the item response theory. In the main study, this selection of item triads was validated by a sample of $N = 55$ music students in a comprehensive test setting to determine correlations with other subskills such as ear training skills, absolute pitch, working memory capacity, or melodic memory. The final path model showed that NESI has a strong relationship with general ear training skills but is only slightly correlated with working memory, spatial orientation and melodic memory.

Thinking in music: An objective measure of notation-evoked sound imagery in musicians

The ability to imagine pictures, sounds or smells, which are not physically present, is common to most humans. Aside from this general cognitive performance, imagery is often deliberately used in learning processes. Both in practice and in performance, musicians imagine how their music should sound while reading notated music/musical text and try to optimally approximate this representation. For example, a violinist might imagine the next played sequence of notes, a jazz trumpeter improvises a melody in the bebop style or a singer anticipates the richness of vibrato for the next phrase.

An overview of historical accounts and qualitative findings of composers concerning the use of imagery was given by Agnew (1922) and of living musicians by Bailes (2009). Already in his *Advices to Young Musicians*, Robert Schumann (1854/2002) recommended to young composers not to try a piece on the instrument until it has been fully conceived (p. 63). Another interesting case is the conductor Dimitri Mitropoulos who was famous for having a clear inner imagery of extremely complex compositions, such as Strauss' *Elektra* or Berg's *Wozzeck* (see Trotter, 1995).

Auditory or sound imagery also plays a more general role in musical memory aside from just in the impressive memory of certain musicians. For example, Highben & Palmer (2004) investigated the influence of two practice types on a memorized piano performance and controlled for auditory imagery by using a same–different comparison task for melodies similar to Wing's *Tests of Musical Ability and Appreciation* (Wing, 1971). In this task, participants heard a melody, read a slightly different or the same melody and had to state whether the heard and read melodies were the same. This study provides evidence for a successful memorization of

the musical scores after a mere auditory or motor practice task. Those participants with a high score on the melody comparison task performed better in the motor condition (lacking auditory feedback) than those with a lower score. However, Wing himself considered this a test of memory while Highben and Palmer interchangeably employed this test to measure auditory imagery and aural skills. Due to the dated test development by Wing, we cannot be sure what this test actually measures.

Apart from its role during practice, musical imagery is also relevant for the unrehearsed performance of music, so-called sight-reading: In studies, notational evoked imagery was the second most important predictor for sight-reading performance – aside from practicing sight-reading directly (Kopiez & Lee, 2008; Kopiez, Weihs, Ligges, & Lee, 2006). In this task, imagery abilities seem to help musicians anticipate the upcoming sound; a match between imagined and heard sound acts as an online validation of the played music without a retrospective comparision.

How can the use of musical imagery as a skill be measured? Although the relevance of musical imagery for professional musicians is clear, an objective measurement of the development of this skill remains a methodological challenge. Probably the most-used and most widely distributed test to measure imagery or a related concept in adults is the *Advanced Measures of Music Audiation* (AMMA) by Gordon (1989). In this test, participants are required to hear two melodies and state whether they are the same or if they are either tonally or rhythmically different.

Gordon also defined two types of audiation as notational audiation: 1) “reading the notation of familiar and unfamiliar tonal patterns and rhythm patterns in familiar and unfamiliar music” and 2) “writing from dictation the notation of familiar and unfamiliar tonal patterns and

rhythm patterns in familiar and unfamiliar music" (1989, p. 12). As more generally defined, "Notational audiation takes place when one hears music seen in notation when the sound is not physically present. One may notationally audiate by reading, writing, or composing music" (Gordon, 1986, p. 13). Undoubtedly, a musician needs to be able to read and write music notation for this task. At the same time, Gordon assumed that the AMMA does not depend on musical expertise and is aimed at an audience of college and university students. However, a recently conducted item analysis of the AMMA revealed that due to its insufficient internal test validity and low task difficulty, this test is not suitable to specifically measure notational audiation, especially not among professional musicians (Platz et al. 2015).

In the 2000s, Brodsky and colleagues (Brodsky, Henik, Rubinstein, & Zorman, 1999; 2003; Brodsky, Kessler, Rubinstein, Ginsborg, & Henik, 2008) developed a new paradigm to test for the development of auditory imagery among the population of professional musicians. In tests using his "embedded melody paradigm," participants read a variation of a short melody from score in silent mode (without singing or humming) for a maximum of 60 s followed by listening either to the underlying original theme or a similar but mismatching theme (a so-called "lure" melody) without seeing the score. Afterwards, participants had to decide whether the heard melody was identical with the notation of the read score. This task could only be accomplished if the participants had developed auditory imagery. Surprisingly, Brodsky et al. (2003) reported that only one third of orchestra musicians were able to perform this task reliably.

Much easier and more accessible to the general population is the task given in the Bucknell Auditory Imagery Scale (BAIS) (Halpern, 2015). The scale consists of 28 self-report items for which participants have to rate based on the vividness of an imagined sound and how well they can imagine the sound changing. Some of the sounds are musical, but most of them do

not contain any musical connection. This scale was used by Gelding, Thompson, & Johnson (2015) during their development of a performance test, the Pitch Imagery Arrow Task (PIAT). This task does not rely on explicit musical knowledge but asks for a vivid representation of a musical scale, in which the participant is supposed to move in steps. At first, the steps up or down are heard and accompanied by up or down arrows. Later, based only on the visual arrows, participants are required to imagine the steps within this tonal context and, finally, they have to judge whether a played tone corresponds to the last imagined tone. The PIAT incorporates a performance task at the end and is therefore a much-needed improvement to measure musical imagery among the general population. In addition, participants have to manipulate musical material, which is a core feature of imagery. Evidence for convergent validity is given by its high correlation with the self-report inventory, the Bucknell Auditory Imagery Scale (BAIS) (Halpern, 2015).

However, the PIAT is not suitable as a substitute for measuring expert musicians' musical imagery skill, because it asks for a more general concept of musical imagery and is probably too easy for expert musicians: The musician group in Gelding (2015) already solved 82% of the trials correctly, but this group consisted of participants with more than five years of training. Thus, the group might have included expert musicians but was also made up of amateurs. If the PIAT was already rather easy in this group, a group of expert musicians would be expected to show ceiling performance.

To investigate how well musicians can imagine the sound of notated music, we adopted the embedded melody task by Brodsky, et al. (2003; 2008). This task is based on producing a sound image of the notated variation in one's mind to compare it afterwards to the heard theme. Moreover, the read melody is usually more complex than the heard melody, which means that

from the point of view of working memory, participants have to actively decompose the complex information into memory units (so-called chunking) and manipulate them in working memory. During this task, participants are not allowed either to produce any accompanying sounds by singing or humming or to use motoric memory by, for instance, silently playing a piano on the table. Such helpful mechanisms were specifically investigated by Brodsky et al. (2003; 2008). The embedded melody paradigm has also worked well in previous studies (Kopiez & Lee, 2006; 2008). However, the original material by Brodsky et al. (2003; 2008) was based on familiar melodies, which probably introduced a bias as recognizing a familiar theme within a notated variation is assumedly easier than if the material is unknown. Whereas unknown material has to be chunked, a familiar theme has already been stored in preprocessed memory units, which only have to be retrieved form long-term memory.

Thus, following this work and our slight adaptations of the paradigm, we have defined the investigated construct as follows: *The use of notation-evoked sound imagery (NESI) is a cognitive skill to abstract relevant deep musical structures from surface musical features by using mental operations without producing corresponding sounds or movements overtly (e.g., by singing or humming). This ability is present when a participant correctly judges the similarity of a notated variation with its musical theme, which may match or mismatch the harmonic structure of the variation.* Our first aim in this study was the development of a test using nonfamiliar musical materials to measure this skill (Pilot Studies 1–3).

Pilot Studies 1 and 2

We conducted Pilot Studies 1 and 2 to develop an item pool with items of various difficulty and to review and improve the rehearsal phase and the instructions to the participants. We conducted

Pilot Study 3 to investigate further properties of the items and produce a model-based statistical test.

Method of Pilot Studies 1 and 2

Stimuli construction. Prior to data collection, two experts from the department of music theory/composition at the Hanover University of Music, Drama and Media developed item triads according to the embedded-melody paradigm (Brodsky et al., 2003). Each item triad consisted of three interconnected melodies: a) a musical theme, b) a variation that harmonically matched this theme and followed the concept of a figural variation, and c) a lure, which was a harmonically mismatching but plausible variation that could be mistaken as a figural variation of the theme (for an example, see Figure 1 and sound examples from the Supplemental Material Online Section, Audio S1). Pilot Study 1 started with a total number of 27 item triads.

********** Place Figure 1 about here **********

Due to the concept of figural variation, variations and lures contained more notes than the respective theme. Thus, each read variation or lure on its own could be harmonically reduced to a set of themes, because a single melody can be accompanied by a number of (implicitly evoked) harmonies. If participants only performed a harmonic analysis of the read score, they would often say that the score did not match the heard theme, because they would have produced a possible harmonic reduction, but not the exact harmonic reduction that was reflected by the theme. Rather, a step by step comparison of the variation (or lure) and the theme has to be performed to find out whether the two presented melodies actually match or not. This can only

be achieved when participants imagine the sound of the read melody while hearing the theme and then compare if the harmonic structure is the same.

All items were reviewed by the authors of this study and then, if necessary, improved by the respective item creator until the variation clearly matched the theme in contrast to the lure. During this process, all parties involved made suggestions to increase the musical meaningfulness of the melodies.

Procedure. In a computer-controlled presentation, each item triad was the source for two possible items: theme plus (matching) variation or theme plus (mismatching) lure. An item was displayed to the participants who, in a first step, looked at the score of either the variation or the lure, were given its starting pitch and tempo from a short audio file, and were then instructed to imagine this melody. In the second step, after participants had indicated that they had imagined this melody as clearly as possible, the score disappeared, and they heard the theme once. In the third step, after comparing the imagined variation/lure and the heard theme, participants entered whether both melodies matched or not. To guarantee that the participants had understood the task, they had to read an extensive explanation of the task and complete three practice trials with feedback in case of wrong answers. This instruction phase was improved on from study to study based on the questions that the participants asked the examiner (for the final version, see Supplemental Text S1).

Design. The design of Pilot Studies 1 and 2 was balanced insofar as each participant always saw either the variation or the lure from one item triad. Participants were randomly allocated to one test condition from a balanced block design. Of the 27 item triads –hence 54 paired items – each participant was given 18 items. Additional questions were asked about their sex, age, main instrument, length of playing time, and self-reported absolute pitch (AP). The

questionnaire was implemented in the online questionnaire provider *SoSci Survey* (https://www.soscisurvey.de), which allowed an individual randomization of the sound imagery items for each participant. Pilot Studies 1 and 2 were conducted at the Hanover University of Music, Drama and Media in a lab setting in small groups (up to six participants).

Ethical Approval. These and all subsequent studies were performed in accordance with relevant national guidelines and regulations (German Psychological Society, 2016). Informed consent was obtained from all participants. Anonymity of participants and confidentiality of their data were ensured.

Pilot Study 1

Participants. Participants in Pilot Study 1 included 43 musicians and music students from the Hanover University of Music, Drama and Media. Their mean age was $M = 30.1$ years ($Mdn = 25$, $SD = 11.0$, 30 men). The main instrument of most participants was the piano (21), followed by flute (4), guitar (4) and voice (3); all other instruments were played by only one participant. Due to a technical difficulty during data collection, one item had to be removed from the analysis. The entire procedure lasted 45 minutes. Participants were financially reimbursed (15 €).

Data Analysis. As we were mainly interested in the item characteristics, data analysis did not focus on the participants, but only on the item triads. The structure of the collected data followed a 2 × 2 matrix design: Each paired item could be a match (variation plus theme) or a mismatch (lure plus theme). Also, participants could state that the read and heard melodies were a match or a mismatch. According to Signal Detection Theory (Macmillan & Creelman, 2004), counting how many participants correctly detect a match in a variation-plus-theme-item within

an item triad produces the number of *Hits*. Counting how many participants wrongly detect a match within a lure-plus-theme-item within the same item triad produces the number of *False Alarms*. The sum of the z-transformed hit and false alarm rates is the basis for the item difficulty by means of sensory/information discrimination, i.e., d' = z(*Hit rate*) – z(*False Alarm rate*) and the item response bias $c = -.5 \times$ [z(*Hit rate*) + z(*False Alarm rate*)].

Results. The mean item difficulty d' of the 27 item triads was $M_{d'} = 1.30$ ($SD_{d'} = 0.84$; for a distribution of d' values, see Supplemental Figure S1). The mean item response bias c of the 27 item triads was $M_c = 0.12$ ($SD_c = 0.31$). The mean response probability measured as percent correct of the variations was $M_{PC\_var} = .69$ ($SD_{PC\_var} = 0.12$) and of the lures $M_{PC\_lur} = .74$ ($SD_{PC\_lur} = 0.17$). A small, non-significant difference between men and women was detected (Cohen's $d = 0.30$; better performance by men).

Discussion. These preliminary results showed that for an item pool with a wide range of difficulties and a subsequently balanced test construction, more difficult items were needed: Only eleven of the 27 item triads showed a sensitivity lower than $d' = 1$, which corresponds to 69% correct detection. If items were already rather easy for music students; experienced musicians, conductors or composers would presumably show ceiling effects and render the test inefficient. More items, which we assumed would be rather difficult, were developed and tested in Pilot Study 2.

Pilot Study 2

Materials. According to the principles for the item construction used in Pilot Study 1, we composed 13 new item triads of higher difficulty than those used in Pilot Study 1. This was

attempted by producing longer, harmonically and melodically more complex items, as well as more subtle differences in the implicit harmonic structure.

Participants. Participants in Pilot Study 2 consisted of 48 music students from the Hanover University of Music, Drama and Media. Their mean age was $M = 23.6$ years ($Mdn = 23$, $SD = 3.2$, 22 men and 26 women). The main instrument of most participants was the piano (11), followed by violin (8), voice (6), trumpet (4), percussion/drum set (4), electric bass (2), clarinet (2), flute (2), saxophone (2); all other instruments were played by only one participant. The entire procedure lasted 45 minutes. Participants were reimbursed (20 €).

Design and Procedure. The same block design as in Pilot Study 1 was used. Of the 13 item triads developed for Pilot Study 2 – hence 26 paired items – each participant was given 13 items in addition to three items from Pilot Study 1. Again, the mix of matches and mismatches was balanced in this item selection.

Participants returned to the lab one week later and were given a set of eight items from the items they had tried to solve at the first session. All participants were given the same eight items. By these means, we were able to analyze the stability of the item characteristics.

Data Analysis. The same procedure as in Pilot Study 1 was used.

Results. The mean item difficulty *d'* of the 12 item triads was $M_{d'} = 0.81$ ($SD_{d'} = 0.45$; for a distribution of *d'* values see Supplemental Figure S2). The mean item response bias *c* of the 12 item triads was $M_c = –0.17$ ($SD_c = 0.30$). The mean response probability in percent correct of the variations was $M_{PC\_var} = .70$ ($SD_{PC\_var} = 0.46$) and of the lures $M_{PC\_lur} = .58$ ($SD_{PC\_lur} = 0.50$). Again, a small, non-significant difference between men and women was detected (Cohen's $d = 0.46$; better performance by men).

The item difficulty as well as the response bias were compared between the test and retest (see Supplemental Figure S3). The correlation of the retest reliability for both characteristics was high ($r_{d'}$ = .79, $t(6) = 3.14$, p = .02 and r_c = .90, $t(6) = 4.99$, p = .002). Therefore, item characteristics could be regarded as very stable.

Discussion. After this second item development and pilot run, 11 rather difficult items ($d' < 1$) were added to the initial item set of Pilot Study 1. Five items of Pilot Study 2 could be described as medium difficult or easy ($d' > 1$). Altogether, with the items from Pilot Study 1, the item set contained enough items of various difficulties to continue with Pilot Study 3.

Pilot Study 3

The purpose of this study was to develop a model-based test to measure the notation-based sound imagery skills of (future) professional musicians based on a complete item presentation.

Method

Item selection from the Pilot Studies. We included three criteria to objectively select item triads from both pilot studies and use them in Pilot Study 3. Generally, the item difficulty of the complete triad had to be above chance level ($d' > 0$). However, we first decided to set the limit even higher to $d' \geq 0.5$ to definitely avoid items where even the highest-skilled participants could only guess. Second, both paired items were supposed to have the same level of difficulty. This criterion was reflected in the response bias, which had to lie within c = [-0.4, 0.4]. A high absolute value of c means that the response probability for lure–theme and variation–theme are rather dissimilar. Third, to complement the human assessment of the harmonic match between variation and theme as well as the mismatch between lure and theme, we included an algorithmic measure of harmonic similarity from the *Simile* package (Müllensiefen & Frieler, 2004; 2007).

The chosen feature was *harmCorE*, which relied on the simple Edit Distance between the two melodies and compared the melodies' implied harmonies. The main critierion was that the harmonic similarity between variation and theme had to be equal to or larger than the harmonic similarity between the lure and theme: *harmCorE*(variation/theme) ≥ *harmCorE*(lure/theme).

By these means, the initial item pool of 39 item triads from Pilot Study 1 and 2 was reduced to 20 item triads that met all the above criteria. In this new, reduced item pool the item difficulty range was within *d'* = [0.52, 2.96], the response bias *c* = [-0.39, 0.35], the response probability for variation and theme *PC_var* = [.52, .94] and for lure and theme *PC_lur* = [.52, .92]. In a randomized and blinded manner, one paired item (10 items with variation/theme and lure/theme, each) was drawn from each item triad and used in Pilot Study 3.

Design. In a similar manner as in Pilot Studies 1 and 2, participants were asked about their sex, age, main instrument, length of playing time, and self-reported absolute pitch (AP). The questionnaire was again implemented in the online survey platform *SoSciSurvey*, which allowed an individual randomization of the sound imagery items for each participant and the complete administration of the study. All 20 items were displayed to all participants. Therefore, for the first time in this project, no data were missing by design, and we obtained a full data set of each participant as well as a much higher test power.

Participants. The data were collected at the universities of music in Lübeck, Munich and Stuttgart (all in Germany). All $N = 135$ participants were music students (73 women, $M_{age} = 23.4$ years, $SD = 4.8$, $Mdn = 22$). Piano, voice, violin and trumpet were most often the main instruments.

Data Analysis. For the construction of a homogenous test to measure a musician's sound imagery skill, we employed statistical tests from item response theory, specifically, the Rasch

model (1PL model) (Bond & Fox, 2007; De Ayala, 2009). The model fit of the items for Rasch model conformity was tested using three measures from the eRm package in R (Glas & Verhelst, 1995; Mair & Hatzinger, 2007; Mair, Hatzinger, Maier, & Rusch, 2016; R Core Team, 2017): (a) the χ^2 itemfit statistic, which is based on the item residuals, namely, the difference between the observed and estimated value of each item, and determines how well the item fits the model using the χ^2 distribution; (b) the *LR-test* (likelihood-ratio test) (Andersen, 1973), which is applied on the item level and investigates whether the remaining items are too dependent on the tested item or whether, on the other hand, the dimensional structure of the item set might be multidimensional; (c) the *Wald-test*, which compares the parameters of item groups or single items and investigates whether systematic deviations between the item parameters, relative to their standard errors, are present.

Significant outcomes in the LR-test and the Wald-test are indicators for items' non-model conformity because of multidimensionality, different item parameters in subgroups (i. e., subgroup invariance) or local stochastic dependence, in which case the mathematical properties of a single item are too dependent on the other items. With regards to the possible subgroup difference, we used two split criteria in the following calculations: Participants were classified into one of two groups based on their achievement in the test (high-achiever vs. low-achiever) and based on their sex. The latter split criterium was based on a stable sex difference (higher achievements by men throughout) in aural skills (Wolf & Kopiez, 2018) and tentative results from the pilot studies (non-significant results, but a difference in Cohen's $d_{PS1} = 0.30$ and $d_{PS2} = 0.46$).

Results

Rasch analysis. After the calculation of the χ^2 itemfit statistic, the LR-test and the Wald-test, three items showed significant deviations from the model in the χ^2 itemfit statistic and the Wald-test, in which the mean achievement was used as a split criterion. When we used the participants' sex as a split criterion, we also found no significant results. The statistical details of the 17 items can be found in Table S1.

Effects (participants). Using the remaining 17 items, which now represented a unidimensional construct, we calculated a performance score for each participant (*d'* score from signal detection theory) and detected the following results: No significant difference was found between the students from different universities ($F(2, 132) = 1.14$, $p = .32$) and between male and female students ($t(115.2) = 0.38$, $p = .70$, $d = 0.07$). The participants' age showed no correlation with their performance ($r = –.02$, $p = .85$). The type of main instrument (harmonic, melodic, rhythmic) did not explain any significant difference ($F(2, 132) = 2.33$, $p = .10$) although the five participants with a percussive main instrument showed a much lower mean performance ($d' = 0.59$) than both groups with a harmonic or melodic main instrument ($d' = 1.57$, $d' = 1.29$, respectively). The duration of lessons on their main instrument was as irrelevant as the participants' age ($r = –.01$, $p = .87$), the latter probably due to too little variance ($SD = 2.50$ years). Only the self-reported AP possessors (n = 8 participants) provided a significant between-groups difference ($t(11.5) = 2.23$, $p = .023$ [one-tailed], Cohen's $d = 0.77$).

In the last step, the number of items was reduced to create a shorter and more economic test instrument for future applications. Twelve paired items (six paired items from the variation/theme and lure/theme subset each) were selected from the sample of 17 items (see also Supplemental Table S1). The six items from the variation/theme and lure/theme subsets each best represented the initial eight variation/theme and initial nine lure/theme items, respectively.

This was quantified by a Pearson correlation coefficient between the mean percent correct of all possible choices of six variation/theme items with the full eight variation/theme items (28 correlation coefficients) and repeated for the best choice of six out of nine lure/theme items (84 correlation coefficients). The highest correlation coefficients for both item groups were $r = .93$ (for reference, the lowest correlation coefficients were $r = .86$ and $r = .87$, respectively). Correspondingly, the sensitivity and bias between the initial 17 and final 12 items also correlated highly with $r_{d'} = .93$ and $r_c = .90$.

The difficulty range of the response probabilities of these final twelve items might seem too narrow [.61, .79] with too few easy items. Nevertheless, the present test measures skills that have been – in an optimal scenario – developed by experts over decades. Moreover, easy items would not discriminate between trained musicians and would fail to add information to a participant's score. An adequate range of item difficulties is also affirmed by the distribution of percent correct scores among participants with $M = 0.69$, $Mdn = 0.75$, a rather large $SD = 0.19$ and by only 10 out of 135 participants who solved all items correctly. To reconnect this test, made up of twelve items, to classical test theory, the separation reliability (Mair et al., 2016) provides a similar measure as Cronbach's α. For the final set of twelve items, SepRel was .57. Admittedly, this value is rather low when benchmarked as Cronbach's α. Several explanations are possible: First, we still do not know enough about the cognitive processing of melodies – and how much this varies between participants. Second, we do not want the *SepRel* to be close to one. If it were, we would measure a very narrow construct, which could be measured by a single item. Last, low reliability coefficients are no exception in musical skills tests: The split-half reliability for the Wing Standardized Tests of Musical Intelligence (Wing, 1961) lies between .65 and .85 for the various subtests (Shuter-Dyson & Gabriel, 1981, p. 280). And even modern tests,

such as the melodic memory task of the Gold-MSI (Müllensiefen, Gingras, Musil, & Stewart, 2014) produced a Cronbach's α = .61, and subscales of the PROMS-S test (Zentner & Strauss, 2017) produced a McDonald's ω = [.57, .79]. Higher reliabilities are produced by tests in which the musical material is much more standardized and less ecologically valid, such as in the Musical Ear Test (Wallentin, Nielsen, Friis-Olivarius, Vuust, & Vuust, 2010).

It should be kept in mind for future studies that items for a parallel test already exist: For instance, as the variation/theme item was validated in Pilot Study 3, the lure/theme item with a similar difficulty could be used after an examination of its features. By means of this item development procedure, the way to a parallel test is already pre-paved.

Discussion. In Pilot Study 3, we investigated a set of twenty items with an identical task (the imagination of a notated melody followed by a comparison of its harmonic structure to a second, played melody) but with different musical materials. As expected, most of the items (17/20) followed a unidimensional structure and therefore measured the same construct. In accordance with performance tests for expert populations, the items are overall rather difficult to discriminate well even between the highest-ranking musical experts. The non-significant, but possibly relevant performance dip of percussionists seems to be self-evident in a test, which focuses on harmonic analyses and excludes rhythm for the most part. This study also served the purpose to investigate the notational imagery skill without a selection bias towards participants. It allows the conclusion that this test is suitable for all music students (and can even be attempted by every person who can read music), while it is perfectly possible and probable that several music students will only perform at chance level.

Up until now, there has been little research on the role of working memory capacity or aural skills on musical imagery tasks in advanced musicians. Bishop, Bailes, & Dean (2013a,

2013b) found no clear correlations between musical imagery and working memory in music production tasks. Nonetheless, we assume that the notation-based NESI skill is dependent on working memory (general capacity as well as a specific musical capacity) and aural skills. Either correlation can only be measured once a well-developed test exists; only then is it possible to measure the NESI skill in an unbiased and scrutinized way.

Main Study

Method

Aims and Design. The main study, which used the NESI test amongst others, followed a descriptive correlational design. In line with our hypothesized correlations of sound imagery with aural skills as well as musical and general working memory, we were first interested in the underlying structure of relationships between various music-specific and general cognitive variables. The relationship between subskills would best be made transparent by a path model. Second, the validity and reliability of the NESI inventory were to be tested.

Participants. Fifty-five music students from the Hanover University of Music, Drama and Media, Germany (30 women, 25 men) participated (M_{age} = 23.1 years, *SD* = 2.78, Min = 19, Max = 31). As could have been expected, their mean degree of musical sophistication was above average, i.e. in the 90th percentile for the German population, with a general factor Gold-MSI score of *M* = 98.0 (*SD* = 8.3, Müllensiefen et al., 2014; Schaal, Bauer, & Müllensiefen, 2014). Based on an objective performance test, eight participants (14.54%) possessed AP (according to the procedure suggested by Deutsch, 2013). Participants played their main instrument for *M* = 13.5 years (*SD* = 4.19) and showed a large variety in instruments. With an average of semesters *M* = 5.53 (*SD* = 3.44), participants in this sample represented all stages in their course of studies.

Stimuli and Tests. The following tests were used in the Main Study (see flowchart in Figure 2):

1. NESI: The *Notation-Evoked Sound Imagery* skill of the participants was measured by an item set of 12 paired items (six paired items each from the variation/theme and lure/theme subset), which were selected from the sample of 17 items as described in Pilot Study 3. An online version of the NESI inventory based on the final 12 item pairs is available online at http://www.thinkinginmusic.com.

2. META: The *Musical Ear Training Assessment* (Wolf, 2016; Wolf & Kopiez, 2018) measured the participants' analytical hearing skills using the shorter 10-item version (for an online version of this test, see http://www.thinkinginmusic.com). For example, the test asked what the highest scale step within a melody was, what exact triad was heard when given the lowest pitch (root and quality), or which basic rhythmic patterns made up a certain rhythm. This assessment was developed using a large-scale online and a longitudinal validation study, in which its unidimensional structure, criterion-related validity and construct validity were corroborated.

3. Gold-MSI and 4. MMT: Aside from the above mentioned 18-item general factor of the *Goldsmiths Musical Sophistication Index* (Müllensiefen et al., 2014), we also assessed the *Melodic Memory Test* in the Rasch-modeled version from Harrison, Musil, & Müllensiefen (2016, Section 5.1.3), who maintained, "The essence of the melodic discrimination paradigm is a similarity comparison task that depends strongly on the limitations of working memory" (p. 3). Thirteen melody pairs were presented to the participants who had to decide whether they were the same or different and then rate how confident they were in their answer (3-step confidence scale). The second melody was always transposed by a semitone or a fifth. We used the area under curve (AUC) score as the outcome variable, which was influenced by the confidence rating.

5. FiguVar: To control for the skill of detecting harmonically appropriate figural variations by reading a score, we excluded the aural presentation of the theme melody from the NESI test and asked participants to solve items by viewing them on the screen. We used eight items that were successfully pretested but did not make it into the final composition of the inventory. The theme and variation were presented one above the other on one screen. As we

assumed this test to be rather easy, each melody pair was only displayed for 15 seconds, after which the response field appeared.

6. BAIS: For a more general perspective on the imagery skills of our participants, the *Bucknell Auditory Imagery Scale* (Halpern, 2015) was tested. A sum score for both the *Vividness* and *Change* subscale was calculated.

7. STM: The *Short-Term Memory* was measured using a digit span forward task by means of a researcher-developed software [Working Memory Central] (Oberauer, Süß, Wilhelm, & Sander, 2007; Oberauer, Süß, Wilhelm, & Wittman, 2003; Sander, 2005). The exact procedure is described in Sander (2005, Section 3.2.10). Numerical sequences from four to nine digits were each tested with increasing difficulty. The digits were presented sequentially, and each sequence length was tested by three items. The score was made up from the overall number of correctly reproduced digits in the correct order.

8. WM: The *Working Memory* was also measured using researcher-developed software [Working Memory Central]. The test consists of a number-based memory updating task using a 3 × 3 matrix with two to seven active cells (Oberauer et al., 2003; 2007; Sander, 2005). In these cells, a digit appears sequentially in each of the active cells as well as an arrowhead to indicate the addition of 1 (arrowhead up) or the subtraction of 1 (arrowhead down). Participants have to remember and indicate the results of these calculations in the active cells. The exact procedure is described in Sander (2005, Section 3.2.11.1). The score was made up from the overall number of correctly reproduced digits in the active cells.

9. MRT: The *Mental Rotation Test* by Vandenberg & Kuse (1978) in the redrawn version by Peters et al. (1995) was employed to measure the mental rotation skill. This skill is one of the few skills that consistently shows large sex differences (Voyer, Voyer, & Bryden, 1995), which

are only partially mediated by differences in spatial working memory (Kaufman, 2007). This test was included so that we could further investigate possible reasons for the observed sex differences (see Pilot Study 1 and 2).

10. AP: Finally, in line with Deutsch's 36-item procedure (Deutsch, 2013), we tested whether the participants possessed absolute pitch. If participants missed the correct pitch by a semitone, their response was counted as correct. Participants were then classified at the 80%-mark (score of 28.8 items correct), which also corresponded to the biggest gap in the score distribution.

The first six measures were tested on the in-browser questionnaire platform *SoSci Survey* (https://www.soscisurvey.de), the memory measures (7 & 8) with the researcher-developed software, and the last two tests (9 & 10) with pen and paper.

********** Place Figure 2 about here **********

Procedure. Participants were tested in small groups of two to five in a lab setting. The whole procedure lasted 2–2.5 hours, and all participants received a reimbursement of 30 €.

Before the data collection started, participants were informed about the study and gave informed consent (see section on ethical approval above). This step and the following were implemented in the platform *SoSci Survey* (https://www.soscisurvey.de). Next, participants were asked about the average amount of time they had spent per week on various musical activities over the last two years as well as which instruments they had played and for how long. Loudness of headphones was adjusted to a comfortable level, which was kept constant for all tests using sound files. Participants started to solve the items of the NESI and META inventories, filled in

their answers on the Gold-MSI general factor, and assessed items as same or different in the MMT (for the instructions for the NESI task see Supplemental Text S1 in the Supplementary Material Online section). Next, participants undertook the FiguVar test, filled in the BAIS, and entered their sex and age. Then they stated, based on a self-assessment, whether they possessed AP and what study program and semester they were enrolled in. After a short break, participants completed the number sequences of the STM and WM tests (Oberauer et al., 2003). Finally, MRT and AP were tested. The examiner thanked the participants for their time and gave them a short debriefing.

Results

As a first step, outliers were removed from the test scores, which applied to NESI (1 outlier), MMT (1), STM (2), and WM (4 outliers), following the procedure by Leys, Ley, Klein, Bernard, & Licata (2013): All scores with more than three median absolute deviations (MAD) from the median were removed. We increased the threshold of 2.5 MAD from (Leys et al., 2013) to 3 MAD as we expected a large variety of skills among the participants. No outliers were found in META, Gold-MSI, BAIS, FiguVar, and MRT.

Factors of influence on NESI and META. We found small but non-significant differences in the sound imagery (NESI) and analytical hearing skill (META) between male and female participants; men achieved higher scores: NESI: $t(47.7) = 1.45$, $p = .15$, Cohen's $d = 0.40$, 95% CI [-0.15, 0.95]; META: $t(44.4) = 0.93$, $p = .36$, $d = 0.26$, 95% CI [-0.29, 0.80]. A more advanced semester correlated positively with a higher aural skill score in both tests but was higher for sound imagery: $r(52) = .29$, $p = .03$, 95% CI [.02, .52]; META: $r(53) = .13$, $p = .35$, 95% CI [-.14, .38]. Both aural skills correlated positively with the duration of the longest–played

harmonic instrument: NESI: $r(50) = .30$, $p = .03$, 95% CI [.03, .53]; META: $r(51) = .25$, $p = .07$, 95% CI [-.02, .49]) but not the longest-played melodic instrument (both $r < .1$). As expected from previous literature (Voyer et al., 1995), men solved more items in the MRT than women: $t(50.3) = 1.90$, $p = .06$, $d = 0.52$, 95% CI [-0.03, 1.07]).

Surprisingly, we found not even a small relationship ($|r| < .1$) between either the BAIS subscales and NESI: Vividness: $r(52) = .08$, $p = .55$, 95% CI [-.19, .34]; Change: $r(52) = -.08$, $p = .58$, 95% CI [-.34, .19]. We found a similarly high correlation between both subscales, $r(53) = .74$, $p < .001$, 95% CI [.59, .84], as in the original study: $r(74) = .74$; $p < .01$. All correlations between the ten tests and inventories can be found in Supplemental Table S2 and Figure S5.

Path model. We included those variables in the path model, which were measured by a performance test, produced a continuous score for the participants and for which clear hypotheses could be formulated. As these latent variables have been developed and validated in previous studies, it was also possible to specify a path model with the variable scores. For a full structural equation model or a larger path model with more variables, many more participants would have been required; with the current sample size ($N = 55$) such a model would not converge. For the theoretically derived model, we followed these theoretical assumptions:

1. NESI is influenced by MMT (Harrison et al., 2016, Sections 3.1 and 3.2) because when imagining a melody, the participant has to encode and to retain the melody in his or her working memory before comparing it with the (heard) theme. This is a prerequisite without which an item cannot be solved.

2. NESI is influenced by WM because a higher general working memory capacity should facilitate the retention of the imagined melody. This is beyond the scope of the music-specific working memory because the first melody is not heard but only read. Visual imagery of the musical score or motor imagery should also be beneficial for this task.

3. NESI and META are influenced by MRT because some kind of "musical mental rotation" might be needed to structure the musical material–whether with sound, as a score or with motoric imagery–in one's mind. A positive prediction would also be an explanation for the difference between the sexes in terms of aural skills: If aural skills partly rely on some kind of rotation skill, they are more difficult for women, who have learned them less often.

4. MMT is influenced by WM because a larger general working memory capacity should be beneficial for a larger music-specific memory capacity. This effect does not merely rely on short-term memory because the second melody (the comparison melody) in the MMT is transposed to a different key (Harrison et al., 2016) section (3.2).

5. MRT is influenced by WM because a larger working memory capacity generally facilitates skills such as mental rotation (Hyun & Luck, 2007; Kaufman, 2007; Lehmann, Quaiser-Pohl, & Jansen, 2014).

6. NESI and META covary because both variables describe the extent to which a person possesses aural skills as both rely on sound–whether imagined or actually heard (Kühn, 1985, p. 17). Moreover, they can be thought of as the same process but in the opposite direction: For analytical hearing, a person hears musical information and is asked to transform this information into notation-based or, more generally, music theoretical material. For sound imagery, notation is given, and a person is asked to transform this written information into imagery.

These six assumptions result in our theory-based model (see Figure S4).

Validation and specification of the path model. The path model (Beaujean, 2014; Steinmetz, 2014) was specified and validated using the R package *lavaan* (Rosseel et al., 2017). As the test scores were manifest and were not estimated as latent variables from the manifest items, we included an estimation of the measurement error for each variable. Using this procedure, we fixed the measurement errors to the internal consistency as measured by Cronbach's α (MMT, WM) or the split-half reliability (MRT). The tests that were developed in our lab (NESI and META) were modeled using the 1PL- or Rasch-model. Their measurement error can be more accurately assessed using, for instance, their separation reliability (*SepRel*), which measures the proportion of person variance that is not due to error (Mair et al., 2016). The measurement errors were the following: *SepRel*(NESI) = .57 (Pilot Study 3), *SepRel*(META) = .72 (this data set), α(MMT) = .68 (Müllensiefen et al., 2014), α(WM) = .83 (Sander, 2005, p. 190), and *r*(MRT) = .79 (Vandenberg & Kuse, 1978). We used this data set to calculate the SepRel for the short version of the aural skills test (META-10) because this was the first study using this exact 10-item test. The variances of all five variables were constrained to 1 before the analysis. Following this standardization, the proportion of error in the variables was obtained by 1–reliability (Keith, 2015, p. 363). Due to missing data after the removal of outliers, the *full information maximum likelihood* estimation was used.

This theoretically derived model produced a converging model, which, however, yielded rather poor fit statistics (see Appendix A in Beaujean, 2014): $\chi^2(3) = 12.7$, $p = .005$; Comparative Fit Index CFI = .322; Root Mean Square Error of Approximation RMSEA = .242; Standardized Root Mean Square Residual SRMR = .117. Following the modification indices, we undertook three theoretically meaningful modifications: First, we removed the path from MMT to NESI (path estimate ≈ 0). Second, we added a path from WM to META; and third, we added a

path from MMT to META. The former change can be justified in that the general working memory capacity is still important, but the specific task in the MMT is too specific and no longer similar to the task in NESI. The two latter changes can be justified in that working memory capacities–general and musical–are more relevant to analytical hearing than previously assumed.

The second and final model, which was slightly adjusted, produced good model fits: $\chi^2(2) = 1.15$, $p = .56$; CFI = 1.00; RMSEA = .000; SRMR = .033 (see Figure 3 and Table 1). The three highest and significant path coefficients were 1) the covariance between NESI and META, 2) the path from MMT to META, and 3) the path from WM to META. As expected, both aural skills (NESI and META) shared a large amount of their variance so that an increase of 1 SD in either variable corresponded to an increase of 0.685 SD in the other variable (Keith, 2015, p. 246). An increase of 1 SD in working memory capacity or musical working memory increased the analytical hearing skill by 0.361 or 0.440 SD, respectively. The final path model including its path weights can be seen in Figure 3.

********** Place Figure 3 about here **********

********** Place Table 1 about here **********

The most relevant deviation from the hypothesized model was the lack of prediction of the sound imagery skill by (musical) working memory as measured by the Melodic Memory Test (MMT) and the working memory test (WM). The direct influence from WM to NESI was only 0.160, the path from MMT to NESI was even removed in the final model. The path coefficients from WM to MMT and MRT were small, but positive, and the latter was in line with results by Kaufman (Kaufman, 2007).

However, the mental rotation skill provided no explanation for the differences between the sexes in terms of their aural skills. The path coefficients were both negative, and the upper positive limit of the 95% confidence interval (CI) was at 0.2. In the most optimistic scenario, a 1 SD increase in mental rotation skill might lead to a 0.2 SD increase in aural skills. However, as most of the CI was in the negative range, this scenario is rather unlikely and further work will probably not corroborate the hypothesis that mental rotation influences aural skills.

Discussion

In summary, this study is the first approach to develop a standardized and validated test for the objective measure of notation-evoked sound imagery. Additionally, the online version of the test guarantees a user-friendly implementation in music cognition research. In this study, we also introduce the first empirical-quantitative model to describe musical hearing in expert musicians. Our approach was guided by the idea that the use of sound imagery can only by investigated with high validity if based on the objective measurement of skills. Thus, we had to develop an item set that allowed us to achieve content and construct validity. This was the aim of three pilot studies and a main study. In Pilot Study 1 and 2, we developed an item pool with item triads which might be appropriate for the testing of notation-evoked sound imagery. Although we also used the embedded melody paradigm for item construction, in contrast to the studies by Brodsky et al. (1999, 2003, 2008), the musical materials in our study were not based on existing collections of classical themes such as the Barlow and Morgenstern (1988) corpus but on specially composed musical material. The advantage of this approach was that it gave us more control over the item triads: For example, we could exclude the confounding effects of

familiarity with the melodies, and the difficulties of item triads could be adjusted deliberately by structural changes of the musical material. Based on a signal detection paradigm, the item difficulty of the initial set of 27 item triads (comprising theme, lure, and variation combined into 54 paired items) was evaluated in Pilot Study 1. The main finding revealed that a mixture of more difficult and easier item triads was required (for the distribution of *d'* values see Figure S1). Due to the high cognitive load for participants caused by a full experimental design, we decided to use an incomplete experimental design with subsets of 18 item pairs presented in a block design to each of the 43 participants. The main aim of Pilot Study 2 was the evaluation of the 13 newly composed item triads with predicted higher but also lower item difficulty. The evaluation of the 26 item pairs by 48 music students in a block design was based on signal detection theory. Additionally, a first analysis of the test-retest reliability of items revealed a high correlation of $r_{tt} = .79$ (retest one week later) so that a high stability of item characteristics could be assumed. A selection of item triads covering a broad range of item difficulties from Pilot Study 1 and 2 was used in the subsequent Pilot Study 3. The main aim of Pilot Study 3 was the development of a model-based test of item characteristics based on a complete randomized experimental design. A reduced item pool of 20 item triads from Pilot Study 1 and 2, covering a broad range of item difficulty ($d' = [0.52, 2.96]$), was selected and presented to a large sample of 135 music students. A subsequent Rasch analysis (1PL model) identified 17 paired items that showed conformity with the Rasch model, revealing a unidimensional construct of "notational sound imagery." Performance scores did not seem to be influenced by various factors, such as main instrument, years of instrumental/vocal lessons, or age. However, those participants with AP outperformed those with relative pitch. In a next step, the number of paired items was reduced to a final set of 12 item pairs. This final set was characterized by a mean of $M = 0.69$ for

correct responses (*SD* = 0.19). This means that the item difficulty was sufficient for discriminating skills in a group of highly expertized participants. Only 10 out of 135 participants reached a score of 100%. Probabilistic analysis of internal consistency of the set of 12 paired items (so-called separation reliability, which can be compared to measures of Cronbach's α), revealed a value of *SepRel* = 0.57. Of course, when compared to the standard benchmarks of Cronbach's α in general psychological tests, this value seems low. However, if compared to values for internal consistency of music-related performance tests, this value seems to be in an acceptable range: Even most advanced tests, such as the Gold-MSI (Müllensiefen et al., 2014), can be characterized by a Cronbach's α of .61. We assume that higher consistency values will depend on a higher degree of standardized musical material, such as those resulting from standardized material or an automatized generation of musical stimuli. For example, the algorithmic and standardized item generation and variation for the melodic memory test (Harrison, Musil, & Müllensiefen, 2016; Harrison, Collins, & Müllensiefen, 2017) reached a Cronbach's α of up to .90. However, automatized item generation would need an elaborated cognitive model of melodic complexity, tonality, and similarity which fully describes the perceptual relationship between theme, variation, and lure. To the best of our knowledge, no such cognitive model for processing complex melodies currently exists for our purposes.

The final aim of the main study was an integration of relationships between subskills into a comprehensive path model. Its theoretical construction considered influences of other factors on NESI performance from previous research, such as music-specific short-term memory, general working memory capacity (WM), spatial orientation, and analytical hearing skills. The regression weights of the resulting model (Figure 3) revealed that there is a strong relationship between analytical hearing skills (as measured by META) and NESI performance, but only a

weak relationship between general working memory capacity and NESI. Our underlying theoretical model assumed that there would be a strong relationship between NESI and WM, but this was not confirmed by the resulting path model. Although the NESI tasks required an involvement of processes usually related to typical WM (e.g., simultaneous storage and processing of information), at first glance, we only found a surprisingly weak direct relationship between WM and NESI ($b = 0.16$, see Figure 3). This result is also in line with Bishop, Bailes, & Dean (2013a, 2013b). However, Ericsson and Delaney's Long-Term Working Memory model for experts (1999) might explain this result. Briefly, Ericsson and Delaney (1999) stated that solving domain-specific tasks generally requires task-specific working memory whose function is to provide highly trainable and thus efficient retrieval processes to access only that information in the long-term memory, which is relevant for the successful execution of the task. Thus, experts' performance depends on the acquisition of knowledge connected with task- specific routines for fast information encoding in long-term memory through retrieval cues as part of their skill acquisition. Since the publication of their study, there have been a number of studies showing low correlations between experts' task-specific performance and skilled everyday activities although the processes might share basic cognitive resources for information retrieval. For example, Chase and Ericsson (1981; Ericsson, Chase, & Faloon, 1980 cited in Ericsson and Delaney, 1999) investigated whether digit span capacity could be improved by deliberate practice. Subjects who successfully practiced their skills to store and retrieve digit sequences showed an increase in digit span but simultaneously a low correlation with span performances in other contents. Practice in digit span leading to superior performance did not affect span performance in other domains. Moreover, low correlations could also be found between tasks of the same domain (Ericsson and Delaney, 1999, p. 264 and 286). The weakness of both

correlation types (tasks between and within domains) could be explained by the maintenance of different task-specific working memory mechanisms and thus different mental representations that are required to solve each task (i.e. digit span ≠ letter span), leading to different transfer of skills depending on the amount of commonality of mental representations (i.e., transfer distance). Related to our study, we therefore conclude that notation-evoked sound imagery is a highly trainable skill connected with task- and domain-specific knowledge. Long-term memory provides specific retrieval processes for an efficient information encoding that is required for the mastery of the measurements' items. Thus, from our point of view, NESI is a skill that can only be acquired and developed by domain-specific deliberate practice in the field of music and not by practice of daily activities (such as mental rotation). To summarize, we think that the low correlation of $r = .089$ between NESI and working memory (WM) together with the medium correlation $r = .38$ between NESI and META should be interpreted as a strength of measurement showing that NESI captures music-related competencies only (in accordance to Weinert, 2001). Future research will have to consider this aspect, and the development of a music-specific WM test should have priority.

Beside the relationships between the measurements used in our main study, we are aware of the superiority of male performances in both music-related tasks, especially for analytical hearing (META) and the use of sound imagery (NESI). However, these differences between male and female achievement in aural skills are only characterized by a small effect size and might be explained by the influence of stereotype threat. In their review of empirical studies, Appel and Kronberger (2012) postulated that academic achievement is not only influenced by subjects' skills and competencies but also by social and situational influences, such as stereotype threats,

which are responsible for an academic achievement gap. Individuals who identified themselves with a negatively stereotyped group showed lower academic achievement compared to individuals with the same skills who identify themselves to be outside of the negatively stereotyped group. According to Steele (1997, cited in Appel and Kronberger, 2012), stereotype threat is experienced by individuals in terms of discomfort especially in evaluative situations resulting in an underachievement for members of the negatively stereotyped group. In detail, stereotype threat describes "an individual's expectation that will adversely influence others' judgments of his or her performance [and] may in turn undermine the individual's actual ability to perform well" (VandenBos, 2007, p. 893). Appel and Kronberger (2012) proposed a three-stage model of stereotype threat in which identification with a negatively connotated stereotype could take place either (1) before learning (2) during learning and its preparation or (3) while taking a test. In the domain of music, especially in the field of music theory and composing, several stereotypes concerning women have been in existence at least since the 19th century. For example, women are not considered to be as good composers as men. Although we do not know which stereotype exactly might have influenced women to underperform in our study, we are aware of the possibility that stereotype threat could have impacted our female participants' achievement. Thus, it will be of future interest to develop procedures for a reduction of potential stereotype influences after having identified them by using an experimental approach.

Another variable of influence is the presence of absolute pitch among expert musicians. In Pilot Study 3, we found that those who reported having AP outperformed those with relative pitch in the NESI task (d = 0.77). This between-groups difference was confirmed in the main study in which the group with AP possessors (n = 8; 14.54%), as classified by an objective pitch identification test, also outperformed the relative pitch group for the correlation between AP and

other variables of the main study see Table S2). This finding contrasts with the results of the study by Brodsky et al. (2008) who found no advantage in the embedded melody task for AP participants. However, the test power of their study was too low to detect a between-groups difference (Experiment 1: n_{noabs} = 16 and n_{abs} = 10) and did not use a performance-based classification procedure. In future applications of analytical hearing tests such as META and NESI, data on AP possessors should be collected as a control variable; as it should be considered for every music performance test focused on pitch (e.g. melodic or pitch discrimination tests).

Finally, only very little is known about the skills of professional musicians to be able to imagine sounds from notation. For example, Brodsky et al. (2008, p. 443) found that only a third of the highly skilled expert musicians that were investigated were able to perform the embedded melody task reliably. This contrasts with Gordon's statement that "unless an instrumentalist or a vocalist can notationally audiate (hear what is seen in notation before it is performed), he or she will be unable to bring meaning to notation as he or she attempts to read it" (1989, p. 12). However, this skill might not be as highly developed—and maybe not as necessary—as Gordon or others might assume. A future aim of ear training in music students could be the development of adequate curricula for the training of sound imagery skills when reading notation. Our suggested test for notation-evoked sound imagery skills can contribute to the evaluation of curricula development, the optimization of learning processes and research into expert musical skills.

Funding information

This work was supported by Pro*Niedersachsen with a grant awarded to the second author.

Supplementary Material

All musical examples of the 12 paired item combinations (Audio S2; 6 pairs of theme/lure and theme/variation each) and the data set used in the path model are available at the Open Science Framework (https://osf.io/hmsy9/ and doi: 10.17605/OSF.IO/HMSY9). An online version of the NESI-test is available online at http://www.thinkinginmusic.com.

Tables and figures/audio files with the index "S" are available as Supplemental Online Material, which can be found attached to the online version of this article at http://??????, under the hyperlink "Supplemental Material."

Acknowledgments

We would like to thank Simon Müller, Nina Düvel and Yves Wyczisk for the preparation of the studies and the data collection in Hanover, Stuttgart, Munich and Lübeck. We are also grateful to Jens Knigge und Adina Mornell for supporting us in the data collection in their departments and to Luis A. Estrada and Arvid Ong for composing the items. Janosch Krämer and Heinz-Martin Süß deserve much thanks for customizing the software Working Memory Central (WMC) to our needs.

The online version of the Musical Ear Training Assessment (META) and the Notation-Evoked Sound Imagery Test (NESI) was created by Sirach Lotz. Both tests are available at http://www.thinkinginmusic.com.

References

Andersen, E. (1973). A goodness of fit for the Rasch model. *Psychometrika*, *38*(1), 123–140. doi:10.1007/BF02291180

Appel, M., & Kronberger, N. (2012). Stereotypes and the achievement gap: Stereotype threat prior to test taking. *Educational Psycholgy Review, 24*(4), 609–635. doi: 10.1007/s10648-012-9200-4

Bailes, F. (2009). Translating the musical image: Case studies of expert musicians. In A. Chan & A. Noble (Eds.), *Sounds in Translation Intersections of Music, Technology and Society* (pp. 41–59).

Barlow, H., & Morgenstern, S. (1988). *A dictionary of musical themes.* London: Faber and Faber. (Original work published 1948).

Beaujean, A. A. (2014). *Latent variable modeling using R: A step-by-step guide*. London: Routledge.

Bishop, L., Bailes, F., & Dean, R. T. (2013a). Musical Expertise and the Ability to Imagine Loudness. *Plos One*, *8*(2), e56052–12. http://doi.org/10.1371/journal.pone.0056052

Bishop, L., Bailes, F., & Dean, R. T. (2013b). Musical imagery and the planning of dynamics and articulation during performance. *Music Perception*, *31*, 97–117.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Brodsky, W., Henik, A., Rubinstein, B.-S., & Zorman, M. (1999). Inner hearing among symphony orchestra musicians: Intersectional differences of string-players versus wind-players. In S. W. Yi (Ed.), *Music, mind, and science* (pp. 370-392). Seoul, Korea: Seoul National University Press.

Brodsky, W., Henik, A., Rubinstein, B.-S., & Zorman, M. (2003). Auditory imagery from musical notation in expert musicians. *Perception & Psychophysics*, *65*(4), 602–612.

Brodsky, W., Kessler, Y., Rubinstein, B.-S., Ginsborg, J., & Henik, A. (2008). The mental representation of music notation: Notational audiation. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(2), 427–445.

De Ayala, R. J. (2009). *Theory and practice of item response theory*. New York: The Guilford Press.

Deutsch, D. (2013). Test for absolute pitch. Retrieved October 13, 2017, from http://deutsch.ucsd.edu/psychology/pages.php?i=6215

Gelding, R. W., Thompson, W. F., & Johnson, B. W. (2015). The pitch imagery arrow task: Effects of musical training, vividness, and mental control. *Plos One*, *10*(3), e0121809. doi: 10.1371/journal.pone.0121809

German Psychological Society [Deutsche Gesellschaft für Psychologie]. (2016). *Berufsethische* Richtlinien [Guidelines for professional ethics]. Retrieved from https://www.dgps.de/fileadmin/documents/Empfehlungen/berufsethische_richtlinien_dgps.pdf

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 69–95). New York: Springer.

Gordon, E. E. (1986). The nature, description, measurement, and evaluation of music aptitudes. Chicago, IL: GIA Publications.

Gordon, E. E. (1989). Manual for the Advanced Measures of Music Audiation. Chicago: GIA Publications.

Halpern, A. R. (2015). Differences in auditory imagery self-report predict neural and behavioral outcomes. *Psychomusicology*, *25*(1), 37–47. doi:10.1037/pmu0000081

Harrison, P. M. C., Musil, J. J., & Müllensiefen, D. (2016). Modelling melodic discrimination tests: Descriptive and explanatory approaches. *Journal of New Music Research*, 1–16. doi:10.1080/09298215.2016.1197953

Harrison, P. M. C., Collins, T., & Müllensiefen, D. (2017). Applying modern psychometric techniques to melodic discrimination testing: Item response theory, computerised adaptive testing, and automatic item generation. *Scientific Reports, 7*(3618). doi:10.1038/s41598-017-03586-z

Highben, Z., & Palmer, C. (2004). Effects of auditory and motor mental practice in memorized piano performance. *Bulletin of the Council for Research in Music Education*, *159*, 58–65. doi:10.2307/40319208

Hyun, J.-S., & Luck, S. J. (2007). Visual working memory as the substrate for mental rotation. *Psychonomic Bulletin & Review*, *14*(1), 154–158. doi:10.3758/BF03194043

Kaufman, S. B. (2007). Sex differences in mental rotation and spatial visualization ability: Can they be accounted for by differences in working memory capacity? *Intelligence*, *35*(3), 211–223. doi:10.1016/j.intell.2006.07.009

Keith, T. Z. (2015). *Multiple regression and beyond: An introduction to multiple regression and structural equation modeling*. New York: Routledge.

Kopiez, R., & Lee, J. I. (2006). Towards a dynamic model of skills involved in sight reading music, *Music Education Research*, *8*(1), 97–120. doi:10.1080/14613800600570785

Kopiez, R., & Lee, J. I. (2008). Towards a general model of skills involved in sight reading music. *Music Education Research*, *10*(1), 41–62. doi:10.1080/14613800701871363

Kopiez, R., Weihs, C., Ligges, U., & Lee, J. I. (2006). Classification of high and low achievers in a music sight-reading task. *Psychology of Music*, *34*(1), 5–26. doi:10.1177/0305735606059102

Kühn, C. (1985). *Gehörbildung im Selbststudium* [Ear training self-taught]. Kassel, Germany: Bärenreiter.

Lehmann, J., Quaiser-Pohl, C., & Jansen, P. (2014). Correlation of motor skill, mental rotation, and working memory in 3- to 6-year-old children. *European Journal of Developmental Psychology*, *11*(5), 560–573. doi:10.1080/17405629.2014.888995

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766. doi:10.1016/j.jesp.2013.03.013

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R, Journal of Statistical Software, *20*(9). doi:10.18637/jss.v020.i09

Mair, P., Hatzinger, R., Maier, M., & Rusch, T. (2016). *Package "eRm."* Retrieved from http://cran.r-project.org/web/packages/eRm/eRm.pdf

Müllensiefen, D., & Frieler, K. (2004). Cognitive adequacy in the measurement of melodic similarity: Algorithmic vs. human judgments. *Computing in Musicology*, *13*(2003), 147–176.

Müllensiefen, D., & Frieler, K. (2007). Modelling experts' notions of melodic similarity. *Musicae Scientiae*, *11*(1 suppl), 183–210. doi:10.1177/102986490701100108

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *Plos One*, *9*(2), e89642. doi:10.1371/journal.pone.0089642

Oberauer, K., Süß, H.-M., Wilhelm, O., & Sander, N. (2007). Individual differences in working memory capacity and reasoning ability. In A. Conway, C. Jarrold, M. Kane, A. Miyake, & J. Towse (Eds.), *Variation in working memory* (pp. 145–215). Cary, UK: Oxford University Press.

Oberauer, K., Süß, H.-M., Wilhelm, O., & Wittman, W. W. (2003). The multiple faces of working memory. *Intelligence*, *31*(2), 167–193. doi:10.1016/S0160-2896(02)00115-0

Peters, M., Laeng, B., Latham, K., Jackson, M., Zaiyouna, R., & Richardson, C. (1995). A redrawn Vandenberg and Kuse Mental Rotations Test: Different versions and factors that affect performance. *Brain and Cognition*, *28*(1), 39–58. doi: 10.1006/brcg.1995.1032

R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org

Rosseel, Y., Oberski, D., Byrnes, J., Vanbrabant, L., Savalei, V., Merkle, E., et al. (2017). *Package "lavaan."* Retrieved from https://cran.r-project.org/web/packages/lavaan/lavaan.pdf

Sander, N. (2005). Inhibitory and executive functions in cognitive psychology: An individual differences approach examining structure and overlap with working memory capacity and intelligence. Aachen, Germany: Shaker Verlag.

Schaal, N. K., Bauer, A.-K. R., & Müllensiefen, D. (2014). Der Gold-MSI: Replikation und Validierung eines Fragebogeninstrumentes zur Messung Musikalischer Erfahrenheit anhand einer deutschen Stichprobe [The Gold-MSI: Replication and validation of a questionnaire instrument for measuring musical sophistication, based on a German sample]. *Musicae Scientiae, 18*, 423–447.

Schumann, R. (1848). *Musikalische Haus- und Lebensregeln* [Advices to young musicians]. Sinzig: Studio.

Shuter-Dyson, R., & Gabriel, C. (1981). *The psychology of musical ability*. London: Methuen.

Steinmetz, H. (2014). *Lineare Strukturgleichungsmodelle. Eine Einführung mit R* [Linear structure equation modelling: An introduction with R]. Rainer Hampp Verlag: München.

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations, a group test of three-dimensional spatial visualization. *Perceptual and Motor Skills*, *47*, 599–604.

VandenBos, G. R. (Ed.) (2007). *APA dictionary of psychology*. Washington, DC: American Psychological Association.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin*, *117*(2), 250–270. doi:10.1037/0033-2909.117.2.250

Wallentin, M., Nielsen, A. H., Friis-Olivarius, M., Vuust, C., & Vuust, P. (2010). The musical ear test: A new reliable test for measuring musical competence. *Learning and Individual Differences*, *20*(3), 188–196. doi: 10.1016/j.lindif.2010.02.004

Weinert, F.E. (2001). Concept of competence: A conceptual clarification. In D.S. Rychen & L.H. Salganik (Eds.), *Defining and selecting key competencies* (pp. 45–65). Göttingen, Germany: Hogrefe.

Wing, H. (1971). *Tests of musical ability and appreciation*. Cambridge, UK: Cambridge University Press.

Wing, H. D. (1961). *Standardized tests of musical intelligence*. Windsor, UK: National Foundation for Educational Research.

Wolf, A. (2016). *"Es hört doch jeder nur, was er versteht." Konstruktion eines kompetenzbasierten Assessments für Gehörbildung* ["For surely everyone only hears what he understands." The Development of a Competence-Based Assessment for Musical Ear Training]. Berlin: wvb.

Wolf, A., & Kopiez, R. (2018). Development and validation of the Musical Ear Training Assessment (META). *Journal of Research in Music Education*. Advance online publication. doi:10.1177/0022429418754845

Zentner, M., & Strauss, H. (2017). Assessing musical ability quickly and objectively: development and validation of the Short-PROMS and the Mini-PROMS. *Annals of the New York Academy of Sciences*, *1400*(1), 33–45. doi:10.1111/nyas.13410

Tables

Table 1. Parameter estimates of the final path model

Path	Estimate	SE	z-value	p-value	95% CI	
NESI~WM	0.160	0.167	0.786	0.432	–0.196	0.458
NESI~MRT	–0.163	0.175	–0.785	0.433	–0.481	0.206
META~ WM	0.361	0.163	2.057	0.040	0.016	0.655
META~MRT	–0.132	0.163	–0.776	0.438	–0.447	0.194
META~MMT	0.440	0.178	2.563	0.010	0.107	0.804
MMT~ WM	0.185	0.169	0.980	0.327	–0.166	0.498
MRT~ WM	0.189	0.177	1.035	0.301	–0.163	0.529
NESI~~META	0.685	0.127	2.594	0.009	0.081	0.579

Note: The path estimates are fully standardized; both the latent and the manifest variables are standardized.

Figures

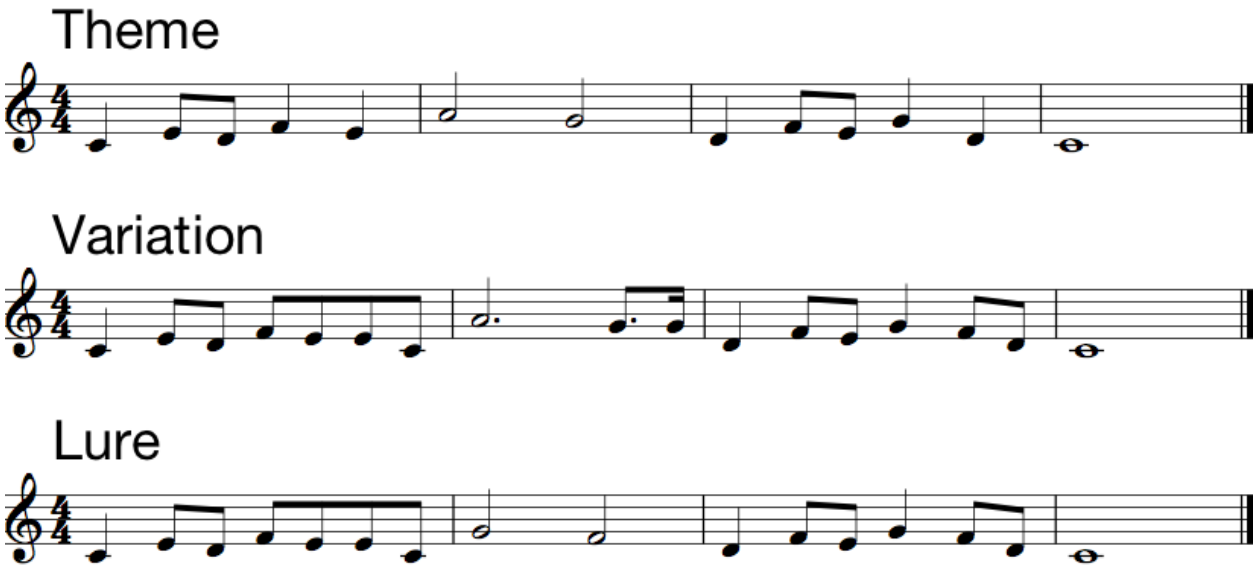

Figure 1. Example of a theme, a harmonically mismatching variation or lure, and a matching variation. See Bar 2 for the most striking harmonic difference (for sound examples, see Audio S1).

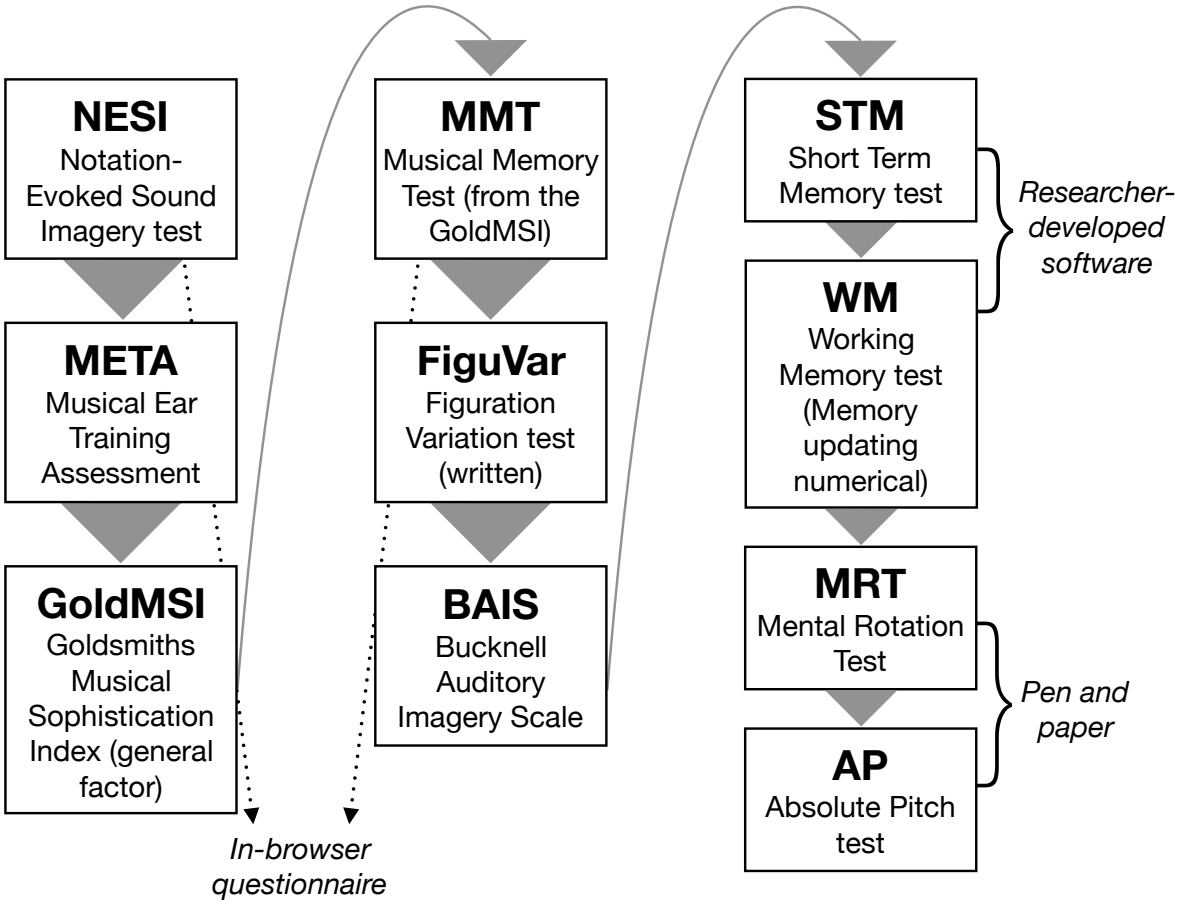

Figure 2. Flowchart of the procedure for the main study (for details, see text).

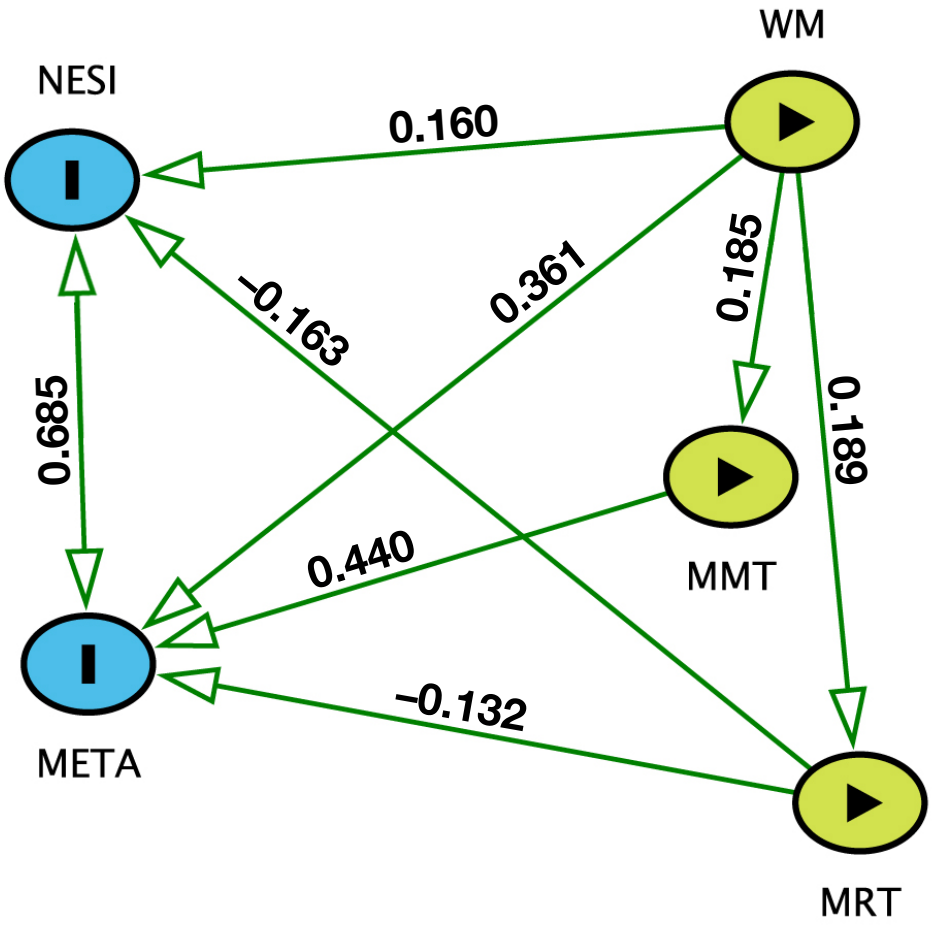

Figure 3. Empirically confirmed path model of the relationship between musical and non-musical components involved in aural skills. The numbers on the arrows are path coefficients. NESI = Notation-Evoked Sound Imagery, META = Musical Ear Training Assessment, MMT = Melodic Memory Test, WM = Working Memory, MRT = Mental Rotation Test.