Measuring Audiation or Tonal Memory? Evaluation of the Discriminant Validity of Edwin E. Gordon’s “Advanced Measures of Music Audiation”

Friedrich Platz¹, Reinhard Kopiez², Andreas C. Lehmann³ and Anna Wolf⁴

Abstract
Edwin E. Gordon developed the Advanced Measures of Music Audiation (AMMA) test to quantify the extent of an adult’s stabilized audiation as a fundamental indicator of musical ability. Although intended to measure audiation exclusively, AMMA is based on a test design similar to the tonal memory subtest of the much older Measures of Musical Talents (SMMT) test developed by Carl Seashore (1919). However, previous studies have shown mixed results regarding AMMA’s construct validity. It therefore remains unclear whether AMMA is suitable for measuring audiation exclusively, as intended by Gordon, or whether it additionally measures tonal memory. Accordingly, we tested this hypothesis in two steps. First, responses of 364 participants were used to identify – in terms of the Rasch model – those items of AMMA that could form a “revised” scale showing measurement invariance; second, we used a Bayesian post hoc correlation analysis (n = 83) to measure the construct (discriminant) validity of the revised version of AMMA compared to an equal number of items in the tonal memory subtest of SMMT. Results from both studies revealed that (a) only five out of 30 items of AMMA showed a model fit that was adequate to form a scale which meets the psychometric requirements of invariant measurement, although with a low internal consistency and an increased probability for ceiling effects, and that (b) both measurements showed a strong correlation (Mdnτ = 0.56, 95% CI [0.42, 0.70], BFₚₒ = 2.67·10¹²). We can thus conclude that there is no practical evidence to assume that both test procedures (AMMA and SMMT) are independent.

Keywords
Advanced measures of music audiation, AMMA, construct validity, Gordon, item response theory, measurement invariance, memory, musical ability, Seashore, tonal memory

Submission date: 11 April 2022; Acceptance date: 7 May 2022

A central question in music education research is how to account for inter-individual differences in music-related achievements (Law & Zentner, 2012; Müllensiefen et al., 2014). With regard to the development of music-related skill, some models still emphasize musical aptitude as an innate variable in (early) childhood. This aptitude is regarded as at least a necessary – if not the most important – factor that may explain and even predict differences between individuals in musical achievement (e.g., Gagné & McPherson, 2019; Schellenberg & Weiss, 2013; Tan et al., 2014). One such approach is Edwin Elias Gordon’s (2012) widely accepted music-learning theory (cf. Shuter-Dyson, 1999, p. 627). In his music-learning theory, Gordon proposes that the upper limit of people’s future music-related achievement is mainly determined by a person’s degree of “audiation”, a core component of innate music aptitude (Gordon, 2012).

¹ University of Music and Performing Arts Stuttgart, Stuttgart, Germany
² Hanover University of Music, Drama and Media, Hanover, Germany
³ University of Music, Würzburg, Germany
⁴ University of Hamburg, Hamburg, Germany

Corresponding author:
Friedrich Platz, University of Music and Performing Arts Stuttgart, Urbanstraße 25, 70182 Stuttgart, Germany.
Email: friedrich.platz@hmdk-stuttgart.de

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access page (https://us.sagepub.com/en-us/nam/open-access-at-sage).
2012). Since its introduction in 1975, and as a neologism combining “audition” and “ideation” (Gerhardstein, 2002, p. 112), the concept of audiation has been revised several times. According to Gordon (2012), “audiation is the process of assimilating and comprehending (not simply rehearing) music momentarily heard performed or heard sometime in the past” (p. 3). Specifically, audiation consists of two interdependent processes, the first of which—the “assimilation” process—is the successful assignment of sensory information to music-related mental patterns regardless of whether this information has been aurally perceived or imagined. The subsequent “comprehension” process of audiation refers to ensuing mental operations including summarizing, generalizing, anticipating, and weighting information based on music-related syntax (Gordon, 2012, pp. 5, 114–116). In the words of Gordon (2012), “sound itself is not music. Sound becomes music through audiation when (...) we translate sounds in our mind and give them meaning” (p. 3). Audiation occurs as a mental activity “by anticipating [emphasis added] in familiar music and predicting [emphasis added] in unfamiliar music what is to come. It involves forward thinking [emphasis added]” (Gordon, 2012, p. 10). Audiation can therefore be characterized as an active cognitive process of creating musical and meaningful units (Gordon, 2012, p. 3). Furthermore, audiation is not a synonym for inner hearing, imitation, memorization, or auditory imagery (Walters, 1989). According to Gordon (2012, pp. 9–12), these skills, unlike audiation, rely on preceding mental events due to their dependence on the successful recall of memorized music without the need to form musical meaning. For example, in contrast to audiation, imitation describes merely a process of copying movements executed on an instrument (or with the voice) without the need for a deeper understanding of one’s own activity (Gordon, 2012, pp. 9, 397).

Gordon’s music-learning theory assumes that music aptitude refers to the innate potential of learning that is reflected in a student’s musical achievements (Gordon, 1989, p. 5, 2012, p. 44), whereby a “window of opportunity” might exist for developing audiation early in life. Once stabilized at around the age of nine, the level of music aptitude developed is presumed to define the upper limit of an individual’s future achievement in music, even in optimal learning conditions (Gordon, 1989, p. 15, 2012, p. 46). Therefore, for a student to achieve a maximum level of musical attainment as determined by his or her music aptitude (that is, stabilized audiation), an optimal learning environment is required in which instructions should, above all other learning preconditions and requirements, be designed to match an individual’s level of audiation (Gordon, 2012, pp. 46–51). To most effectively identify a student’s audition abilities, music educators should call for a valid measurement tool that meets all criteria governing psychometric standards in the related fields of psychological and educational diagnostics (American Educational Research Association et al., 2014). Moreover, the availability of objective and valid tests is in any case a crucial cornerstone for the development of scientific theories (Watson, 2012).

For this purpose, Gordon developed several music aptitude tests intended to measure the music audition of individuals from different age groups from pre-school to adolescence (Gordon, 1965, 1979 & 1986, 1989). These tests have been widely used for assessing music aptitude in various types of school (Moore, 1995) as well as informing research in the fields of music psychology (e.g., Bugos et al., 2007; Burgoyne et al., 2019; Hayward, 2009), music education (e.g., Degé et al., 2017; Schleuter, 1993), and neuroscience (e.g., Schneider et al., 2002; Schneider et al., 2005). In particular, the Advanced Measures of Music Audiation (AMMA) test procedure (Gordon, 1989) has been used as a reference for cross-validating new inventory tests (Law & Zentner, 2012; Müllensiefen et al., 2014; Wallentin et al., 2010). The AMMA can therefore be regarded as a gold standard for music psychology research with young adults and in the development of diagnostic procedures for operationalizing musical aptitude as a latent trait.

The AMMA is intended to measure the degrees of stabilization in the audition of college students, high school pupils, and junior-high students (Gordon, 1989, p. 15). In this test, participants listen to a pair of melodies for every item and must decide whether the two melodies differ due to a possible tonal or a rhythmic change in the second melody. All melodies (and their paired variations) are especially composed, mostly based on modal or atonal-harmonic compositional principles, and are often combined with an irregular rhythm or a ternary metrical structure. Participants have to choose between three responses to the second melody of each pair (s = same; t = tonal change; r = rhythm change) while leaving items unanswered in the case of doubt to minimize guessing as a possible confounding variable (Gordon, 1989, p. 23). The AMMA test consists of 30 items, each item being a melody pair. Among the 30 pairs, there are 10 pairs with tonal changes (so-called “tonal” items), 10 pairs with rhythmic changes (“rhythm” items), and 10 pairs with no changes at all (“same” items).

The development of the AMMA was based on two key assumptions. First, once it has been stabilized, audiation is a latent trait like any other aptitude and is thus normally distributed among the population (Gordon, 1989, p. 10). Second, it is solely responsible for a student’s response behavior on all test items because “a student is expected to audiate concurrently [emphasis added] tonality, keyality, implied harmony, rhythm, meter, and tempo in a test question” (Gordon, 1989, p. 16). The AMMA test thus differs significantly from Gordon’s test procedures for younger children and adolescents, such as the PMMA and the IMMA (Gordon, 1979 & 1986), in which these musical parameters are assumed to be processed independently from each other and are consequently assessed by separate subtests. In essence, the AMMA’s overall test score captures the concurrent feature processing of time- and pitch-dependent auditory information along with the
music-related mental operations relating to stabilized audiation. Most importantly, this characteristic also implies a theoretically desired unidimensionality of the measured latent trait (Gustafsson & Åberg-Bengtsson, 2010, p. 97), and thus offers more clarity in interpreting test scores (Bond, Yan, & Heene, 2021, pp. 31–35).

Intriguingly, 70 years earlier, Seashore (1919) had used a similar approach for constructing his test items which were designed to measure the ability of memorizing tone sequences as a function of tonal memory. In his test of tonal memory, participants were asked to listen to 30 melody pairs consisting of short tone sequences of varying lengths (10 melody pairs each as a three-tone, four-tone, or five-tone sequence). While listening to the second tone sequence within a melody pair, they were asked to identify the specific tone of the second melody that had been altered compared to the initial sequence. Thus, participants had to memorize every tone of the initial sequence as a prerequisite for a successful tone-by-tone comparison while listening to the subsequent tonal sequence. Each participant’s tonal memory span could then be deduced from the number of successfully solved items (Seashore, 1919, pp. 238–243).

To summarize, although Seashore and Gordon intended to measure different and mutually exclusive latent abilities, they used the same construction principle in developing their measurements, namely a melody comparison task in combination with a same/different item response format. To resolve this theoretical conundrum, Gordon decided to separate both melodies within each melody pair by a period of four seconds of silence, which was “found to be optimal for a student to be able to audiate, but not to imitate or memorize the musical question [i.e., the first melody], before he or she hears the musical answer [i.e., the second melody]” (Gordon, 1989, p. 19). However, to the best of our knowledge the question is still unanswered regarding whether the length of silence chosen (i.e., four seconds) between the two melodies is sufficient to isolate or separate the phenomenon of audiation from that of memorization.

The issue of whether the AMMA is an aptitude test that solely measures adults’ music aptitude as a one-dimensional latent variable depends on how far—and with what degree of accuracy—one can infer that audiation is a latent trait based on the test score derived from the measurement (cf. Furr, 2018, pp. 9–11). With regard to the AMMA test procedure we need to take into account the psychometric dimension of invariant measurement (Engelhard, 2013), along with standard psychometric criteria such as reliability (Shrout & Lane, 2012) and validity (Grimm & Widaman, 2012). The procedure hence depends on the degree to which the test’s items are sensitive exclusively to a person’s level of audiation as a one-dimensional latent variable rather than to other possibly latent traits such as (short-term) tonal memory.

Although the validation of test instruments serves multiple objectives (American Educational Research Association et al., 2014), most studies seeking to validate Gordon’s music aptitude tests during the last decades have focused primarily on criterion validity as one indicator of construct validity (Schellenberg & Weiss, 2013, p. 500). In these studies, the degree to which audiation-derived test scores were associated with other music-related achievements that were thought to be predicted by music aptitude was determined by measuring either all variables at the same time (‘concurrent validity’) or in a repeated-measures longitudinal design (‘predictive validity’). In his literature review investigating the criterion validity of Gordon’s audiation tests, Hanson (2019) reanalyzed and aggregated the results of studies in the field of music education research that spanned nearly 50 years. He found a mean correlation (weighted by sample size) of only $r = 0.33$ (95% CI $[0.26, 0.40]$), corrected for artifacts between participants’ composite AMMA scores and all music-related criterion variables from 25 correlations based on a total of 1,063 participants. This correlation of medium effect size (Cohen, 1988) is far from being as strong as would be expected in Gordon’s music-learning theory (2012). We would expect (or hope for) a much higher proportion of shared variance in music-related achievement criteria than the observed 11% explained or predicted by music aptitude as measured with the AMMA test. Consequently, Hanson’s sobering findings suggest that the criterion validity of the AMMA is relatively weak. This finding also contradicts Gordon (1990, p. 11), who reported a much higher predictive validity coefficient in his validation study ($r_{\text{obs}} = 0.82$, 95% CI $[0.66, 0.98]$) (see also Table S1 in the Supplementary Material section). However, Hanson’s results are in line with findings from other validation studies in the field of (music) psychology research, reporting observable correlations as estimators for the criterion validity of the AMMA ranging from $r_{\text{obs}} = 0.02$, 95% CI $[-0.25, 0.29]$ for written harmony (McCrystal, 1995, pp. 33, 35) to a maximum of no more than $r_{\text{obs}} = 0.52$, 95% CI $[0.07, 0.97]$ for students’ achievement scores in music performance (Miceli, 1998, pp. v, 57, 59, 61; see also Table S1; 95% CI recalculated by the authors).

There is a considerable range in estimates of AMMA’s construct validity (Hanson, 2019), with a high proportion of variability due to between-study variability rather than (within-study) sampling error ($I^2 = 69.36\%$). This heterogeneity arises as a result of the difference in the effect size estimators summarized by Hanson, indicating a marked inconsistency in the estimation of AMMA’s validity. For example, Miceli (1998) found a large correlation, albeit with a wide confidence interval ($r_{\text{obs}} = 0.52$, 95% CI $[0.07, 0.97]$; 95% CI recalculated by the authors), and thus only a vague estimation for the association between the composite score of AMMA and the musical achievement score. As she showed, the strength of the association between both scores depended heavily on whether statistical outliers ($n = 4$) were removed from the entire sample of 15 participants or not. She also discovered that music aptitude could not significantly predict more than 3.26% of
the entire sample’s variance in music achievement (Miceli, 1998, p. 62). Indeed, after controlling for statistical outliers, a significant improvement was predicted for music achievement gain scores through composite scores of the AMMA, though only for students with low music aptitude who showed a positive gain in achievement. These varying estimates of predictive validity may partly be explained by the assumption of a sample-dependent sensitivity of AMMA as shown by McCrystal (1995, pp. 34–36), who found generally stronger validity coefficients for participants who showed particularly high or low achievement, but only small correlations between stabilized audiation and achievement grades in music for the entire sample of college undergraduate music majors.

Another reason for this heterogeneity in the estimation of predictive validity might be the large variety of criterion variables. For example, music-related achievements in academic music courses do not usually share more than 50% of common variance with AMMA scores. For example, in his four-year longitudinal study Sang (1998) found the most highly significant raw correlations between the total score on the AMMA and other music-related achievement tests such as ear training ($r = 0.66$, 95% CI [0.31, 0.85]), music theory ($r = 0.51$, 95% CI [0.09, 0.78]), and total music score ($r = 0.48$, 95% CI [0.05, 0.76]). However, due to the small sample size ($N=20$ valid cases), the reported correlations were characterized by large confidence intervals (recalculated by the authors) with lower limits close to Zero. Except for the first correlation (ear training), two further Bayesian post hoc summary analyses for the other two aforementioned correlations showed that there was only “moderate” evidence, according to Jeffrey’s benchmarks for the Bayes factor (Jeffreys, 1961), in favor of the alternative hypothesis (compared to the null hypothesis) for a positive correlation between AMMA and music-related achievements in music theory score ($BF_{1:0} = 6.421$) or in the total music score ($BF_{1:0} = 4.624$). Thus, due to the ambiguous findings regarding its predictive validity, AMMA is not a stable and clear predictor for the overall development of musical achievement, but rather contributes some additional variance to the development of specific kinds of music-related activities (Harrison, 1996).

Heterogeneity was also observed between studies investigating the convergent validity of AMMA (see Table S2 in the Supplementary Material section). For example, whereas Gordon (1989, pp. 49–51) reported strong correlations between the composite score of AMMA and the composite score of the Musical Aptitude Profile (MAP) for undergraduate music majors ($0.58 < r_{obs} < 0.78$), Degé et al. (2017) found only a weak correlation ($r_{obs} = 0.20$, 95% CI [0.06, 0.34]) between 89 children’s test scores on the Intermediate Measures of Music Audiation (IMMA) test and AMMA.

In brief, according to Hanson (2019), the question remains open as to why the criterion validity of AMMA seems relatively low and its validity estimation appears heterogeneous and thus imprecise – even after controlling for study-related factors (by using meta-regression for subgroup analyses) that might be a reason for the observed heterogeneity. Whereas the majority of studies has focused on the external validity of AMMA by investigating associations between the test score and other variables not directly covered by AMMA, only a few studies have concentrated on another facet of construct validity, which is the internal test validity of the AMMA. Generally, in these studies the degree of fit has been investigated between test takers’ observed response behavior against the expected one, ideally predicted by a test theory model (cf. Grimm & Widaman, 2012, p. 623). One of the earliest studies focusing on AMMA's internal test validity is Gordon’s investigation of the latent-factor structure of AMMA for which he used a principal axis factoring approach (Gordon, 1991, pp. 8–21). Based on 5,336 responses from undergraduate and graduate music major and non-music major students, he extracted nine unrotated factors accounting for only 44% of the variance. The rotated factor matrix, however, showed at least one stable factor with an eigenvalue >1, for which only nine out of the 30 items displayed loadings of $\lambda \geq 0.3$, accounting for no more than 11.2% of the variance. In a recent study, Verdis and Sotiriou (2018) reported similarly mixed results: based on a linear factor analysis, they reported up to two uncorrelated factors for AMMA. A subsequent Rasch analysis, revealing a proportion of only 18.1% of the total variance, confirmed the unclear results of their earlier factor-analytical approach. Moreover, their analyses of the residual variances suggest further evidence in disagreement with the single-latent-factor assumption.

Therefore, in order to conclusively affirm that AMMA is suitable for solely measuring audiation exclusive of memorizing as a function of tonal memory, we first have to identify those items in our investigation that can best be assigned to a single, one-dimensional joint variable (of persons and items) as a key requirement of an invariant measurement (Engelhard, 2013, p. 14) while removing all other items that contribute to other latent factors resulting in a significant model misfit. Following Verdis and Sotiriou (2018), we also propose the Rasch methodology as a statistical approach to model participants’ observable responses as a function of their latent trait and of items’ attributes (Bond et al., 2021). The Rasch model explains and predicts the probability of participants’ observable response behavior for dichotomous items as a result of the difference between participants’ ability and items’ difficulties (Bond et al., 2021). Applying this logic to AMMA, the probability of a participant’s $j$ correct response on a dichotomous item $i$ of the AMMA ($y_{ij} = 1$) should only depend on the extent both of the participant’s stabilized audiation ($\theta_j$) and of the item’s difficulty ($\alpha_i$). Consequently, the probability of a participant’s $j$ correct response ($y_{ij} = 1$) on an item $i$ given the participant’s audiation ($\theta_j$) and the item’s
difficulty ($\sigma_i$) is defined in terms of the Rasch model (Hayes & Embretson, 2012, p. 173) as:

$$P(y_{ij} = 1|\theta_j, \sigma_i) = \frac{e^{(\theta_j - \sigma_i)}}{1 + e^{(\theta_j - \sigma_i)}}$$

(1)

The greater a person’s ability, the more items of a scale should be answered correctly, and the higher their probability of success in solving items of a scale showing Rasch conformity compared to the results from people with a lower ability, due to the model’s underlying “non-crossing person response function” (Engelhard, 2013, p. 14). Thus, it is less important which particular items participants answer correctly, but rather the total number of correctly answered items with different levels of difficulty. This reflects a further requirement of invariant measurement, the so-called “item-invariant measurement of persons” (Engelhard, 2013, p. 14). These requirements not only apply to person measurement but also to item calibration (i.e., “person-invariant calibration of test items” and the “non-crossing item response function”, Engelhard, 2013, p. 14). In keeping with the criterion of measurement invariance, scales only fulfill the requirements for valid measurement of person and item characteristics if these items (and persons) show Rasch model conformity. Only then can comparisons between people’s ability characteristics (such as audition) as well as item characteristics (such as difficulty) become possible. Another advantage of a Rasch scale is its use in computerized adaptive testing (CAT), which has been developed and tested, for instance, for measuring music listening skills (Colwell, 2002; Vispoel, 1993; Vispoel et al., 1997), melodic discrimination (Harrison et al., 2017), beat perception (Harrison & Müllensiefen, 2018), and auditory imagery (Gelding et al., 2020; Wolf et al., 2018; Wolf & Kopiez, 2018). In general, adaptive procedures achieve a greater test efficiency with fewer items, a higher level of validity, and a reliability which is similar to that of full-scale measurements.

In terms of scale construction, following Rasch’s methodology and based on the premise of only one “true” latent variable, it is not our objective to find a statistical model that best describes the data, as would generally be the case in social science (“model fits data approach”), but rather to identify those items that could be calibrated on a Rasch scale without showing significant misfit (“data fits model approach”). For determining invariance of measurement as an element of construct validity there exist goodness-of-fit statistics for scales and items that vary with regard to their statistical power (Debelak, 2018; Stone & Zhang, 2003). These include parametric tests of Infit and Outfit (Wright & Masters, 1982, p. 100), Andersen’s Likelihood Ratio test (1973), the Wald test (1943), the Martin-Löf test (1973), bootstrapping tests (e.g., von Davier, 1997), and non-parametric tests (Ponocny, 2001). By applying these tests, it is possible to identify items that do not reflect the same ability and which are thus responsible for the increase in a model’s misfit. We can then exclude these from the resulting final scale. This strict method of item exclusion results in a scale with a high degree of measurement invariance.

Thus, to investigate whether AMMA is suitable for measuring exclusively audition and not concurrently the ability of memorizing tone sequences as a function of tonal memory, we followed a two-step approach. First, to arrive at a revised version of AMMA with a high degree of unidimensionality, we pursued a strategy of identifying through Rasch methodology those items that best fit Gordon’s (1989) theoretical model (Study 1). Second, we used a correlational design to identify the discriminant validity of the revised version of the AMMA compared to the tonal memory test from Seashore’s Measures of Musical Talent (Butsch & Fischer, 1966; Seashore, 1919). The ultimate goal here was to determine the degree to which both tests correlated due to their similarity in item construction (Study 2). In line with Gordon’s model we expect a null correlation ($H_0: \rho = 0$) if the revised version of AMMA reflected purely a person’s music ability (i.e., audition) and not his or her ability of memorizing music (i.e., tonal memory). However, if we find that both test scores correlate, then this association would speak in favor of the concurrent measurement of both latent abilities – and hence lack of independence between them.

**Study I**

**Method**

**Sample.** A total of $N = 364$ participants ($n = 175$ female [48%]) with a mean age of 18.72 years ($SD = 2.12$) took part in our study. To achieve performance characteristics that would be comparable to the norm groups reported in Gordon (1989, p. 38), we selected only students from higher education programs. Specifically, our sample consisted of two subsamples comprising $n = 245$ students of the last two classes of secondary schools (“Gymnasien”, ISCED level 3) in Germany and $n = 119$ undergraduate/graduate music majors from a school of music, respectively (see Table 1).

Although the subsamples’ age and musical background differed, Table 2 reveals equal performances of our subsamples compared to their norm groups (Gordon, 1989, p. 44): secondary school students did not show any significant differences by means of one-sample $t$-tests corrected for six multiple comparisons ($\alpha < 0.05/6$ for each comparison) in their tonal score, $t(244) = -0.44$, $p = 0.66$ (2-tailed), n. s., $d = 0.03$, rhythm score, $t(244) = 2.41$, $p = 0.02$ (2-tailed), n. s., $d = 0.15$, or total adjusted score compared to Gordon’s norm group of high school students, $t(244) = 1.01$, $p = 0.32$, (2-tailed), n. s., $d = 0.06$. Furthermore, no significant differences in the tonal score, $t(118) = 2.10$, $p = 0.04$, n. s., $d = 0.20$, rhythm score, $t(118) = 1.08$, $p = 0.28$, n. s., $d = 0.10$, and total adjusted score were found between our subsample of undergraduate and graduate music majors and their respective norm groups, $t(118) = 1.74$, $p = 0.09$, n. s., $d = 0.16$. Therefore, our total sample best represented the response behavior of the norm groups as reported in Gordon (1989) and comprised
participants from the late-adolescent population who were engaged in cultural activities.

Procedure. Data were collected in small groups with a group size of $n \leq 30$ participants in several secondary schools and universities of music in Germany. Each session lasted no longer than 45 min. First, all participants were welcomed. Next, they were informed by means of a cover story that they would take part in a test to see how musical they were. After giving written informed consent, participants listened to German translations of AMMA test instructions from a CD (Gordon, 1989, p. 23) (see Table S3 in the

Table 1. Summary of total sample’s demographic characteristics and musical background as a function of educational level (Study I).

| Variable                        | Total (N = 364) | Pupils (n = 245) | Students (n = 119) |
|---------------------------------|-----------------|------------------|-------------------|
|                                 | $n$ | $f$ | Min. | Max. | M (SD) | $n$ | $f$ | Min. | Max. | M (SD) | $n$ | $f$ | Min. | Max. | M (SD) |
| Demographic characteristics     |     |     |      |      |        |     |     |      |      |        |     |     |      |      |        |
| Age                             | 361 | 16  | 16   | 34   | 18.72  | 243 | 16  | 16   | 20   | 17.70  | 0.81 | 118 | 18  | 34   | 20.83 | 2.41 |
| Female                          | 361 | 175 |      |      |        | 243 | 114 |      |      |        |      | 118 | 61  |      |      |        |
| Male                            | 361 | 186 |      |      |        | 243 | 129 |      |      |        |      | 118 | 57  |      |      |        |
| Musical background              |     |     |      |      |        |     |     |      |      |        |     |     |      |      |        |
| Age of sustained musical activity | 337 | 3   | 20   |      | 9.34  | (4.50) | 219 | 3   | 20   |      | 9.35  | (4.71) | 118 | 3   | 19   |      | 9.31  | (4.09) |
| Years of private lessons        | 362 | 0   | 22   |      | 7.15  | (4.83) | 244 | 0   | 16   |      | 5.27  | (4.25) | 118 | 0   | 22   |      | 11.03 | (3.44) |
| Years of regular practice       | 360 | 0   | 17   |      | 4.75  | (4.74) | 242 | 0   | 16   |      | 3.35  | (4.18) | 118 | 0   | 17   |      | 7.61  | (4.74) |
| Current time spent practicing    |     |     |      |      |        |     |     |      |      |        |     |     |      |      |        |
| Rarely or never                 | 362 | 91  |      |      |        | 245 | 90  |      |      |        |      | 117 | 1   |      |      |        |
| About 1 h per month             | 362 | 17  |      |      |        | 245 | 15  |      |      |        |      | 117 | 2   |      |      |        |
| About 1 h per week              | 362 | 57  |      |      |        | 245 | 51  |      |      |        |      | 117 | 6   |      |      |        |
| About 15 min per day            | 362 | 49  |      |      |        | 245 | 44  |      |      |        |      | 117 | 5   |      |      |        |
| About 1 h per day               | 362 | 68  |      |      |        | 245 | 36  |      |      |        |      | 117 | 32  |      |      |        |
| More than 2 h per day           | 362 | 80  |      |      |        | 245 | 9   |      |      |        |      | 117 | 71  |      |      |        |
| Concert attendance              |     |     |      |      |        |     |     |      |      |        |     |     |      |      |        |
| (past 12 months)                |     |     |      |      |        |     |     |      |      |        |     |     |      |      |        |
| None                            | 364 | 45  |      |      |        | 245 | 44  |      |      |        |      | 119 | 1   |      |      |        |
| 1–4                            | 364 | 162 |      |      |        | 245 | 135 |      |      |        |      | 119 | 27  |      |      |        |
| 5–8                            | 364 | 65  |      |      |        | 245 | 31  |      |      |        |      | 119 | 34  |      |      |        |
| 9–12                           | 364 | 35  |      |      |        | 245 | 19  |      |      |        |      | 119 | 16  |      |      |        |
| 13 or more                     | 364 | 57  |      |      |        | 245 | 16  |      |      |        |      | 119 | 41  |      |      |        |

Note. $n =$ valid responses, $f =$ absolute frequency. Secondary school students were only selected from the last two grades of secondary schools ("Gymnasium"; ISCED Level 3); items reporting participants’ musical background were obtained from Ollen (2006, pp. 237–239).

Table 2. Summary of test performance as a function of educational level compared to norm groups.

|                                | Sample of validation study (N = 364) | Norm groups for the advanced measures of music audiation |
|--------------------------------|-------------------------------------|--------------------------------------------------------|
|                                | Pupils (n = 245) | Students (n = 119) | High school students (n = 872) | Undergraduate and graduate music majors (n = 3206) |
| Test score                     | $M$ | $SD$ | $M$ | $SD$ | $M$ | $SD$ | $M$ | $SD$ | $M$ | $SD$ |
| Tonal                          | 23.69 | 4.10 | 29.05 | 3.91 | 23.80 | 4.37 | 28.30 | 4.12 |
| Rhythm                         | 27.38 | 3.76 | 31.11 | 3.12 | 26.80 | 4.03 | 30.80 | 3.52 |
| Total                          | 51.07 | 7.24 | 60.16 | 6.66 | 50.60 | 7.91 | 59.10 | 7.41 |

Note. No significant differences were found between test sample and norm groups (corrected significance criterion for multiple comparisons: $\alpha < .008$). The data for the norm groups are obtained from Table 5 in Gordon (1989, p. 44).
Supplementary Material section). Next, participants were requested to provide further information about their age, gender, and musical background on a self-report questionnaire on musical sophistication (Ollen, 2006). In the debriefing, participants were informed of the aims of the study. Later, they received their test scores along with information that interpreted the scores.

Results
The data analysis was performed in three steps, starting with (a) an analysis of response data as well as total scores using classical test theory, followed by (b) a person and item parameter estimation based on the Rasch model (Bond et al., 2021). Finally, to obtain a Rasch-conforming test version of the AMMA, we conducted (c) a test optimization procedure in which model-fit reducing items were removed from AMMA until a sufficient model fit was reached.

In detail, we first prepared the sample’s raw response data for our data analysis by calculating all scores of the subscales as well as the adjusted total score according to the AMMA manual’s scoring procedure (Gordon, 1989). Following Gordon’s (1991) proposed procedure, we also transformed the raw response data from three different response categories for the 30 items (1 = “same”; 2 = “tonal”; 3 = “rhythm”) into a binary matrix, reflecting each participant’s response behavior on every item of the AMMA manifested either as an incorrect (0) or correct (1) solution as the only two and mutually exclusive solution possibilities.

We then conducted descriptive item and scale analyses of the AMMA test using JASP (Version 0.13.1; JASP Team, 2020). We computed all item parameters such as item difficulty, item variance, corrected item-scale correlation, and internal consistency (see Tables S4 and S5 in the Supplementary Material section) as well as all scale parameters in terms of score means, standard deviations, and scales’ intercorrelations (Table S6). Next, we investigated not only the sample’s achievement reflected by the subscales’ scores and the total score of the AMMA (Table S6) but also the participants’ test performances as a function of education level (Table S7) for the purpose of comparing them with their corresponding reference groups (Gordon, 1989, p. 44).

From this point on we used the statistical platform R (Version 4.0.2; R Core Team, 2020) and the eRm package (Version 1.0-1; Mair et al., 2020) to calculate item and person parameter estimates of the Rasch model. First, we included all 30 AMMA items in the initial model (Table 3).

To test whether the initial Rasch model was suitable as a valid formalization of the sample’s response data, we conducted several parametric goodness-of-fit tests, such as the Andersen goodness-of-fit test (Andersen, 1973), the Wald test (Wald, 1943), and the Infit mean-square test (Wright & Masters, 1982) as implemented in the eRm package (Mair et al., 2020). The test results were significant, which indicated that the first model did not meet all criteria of the Rasch model (Table 3). We subsequently performed a model optimization approach in order to arrive at a revised test version of AMMA showing measurement invariance by removing those items that were responsible for the reduction of construct validity (and thus for the significant results from the goodness-of-fit tests). For this purpose, we used Mair et al.’s (2020) optimization strategy, an iterative stepwise item selection procedure whereby the item with the strongest misfit – and thus most responsible for the significant reduction of test validity – is removed in each step until model fit has been attained (Bond et al., 2021; Koller et al., 2012). Rasch conformity of a measurement associated with its uniqueness of unidimensionality and local-specific objectivity (Bond et al., 2021) should thus be achieved with the remaining items. By the end of this iterative, algorithmic approach only five of the original 30 items remained in the revised scale, now optimized in terms of Rasch homogeneity (see Table 4).

We used a more conservative strategy as recommended by Koller et al. (2012, p. 162), because of problematic multiple testing and the low statistical power of the parametric goodness-of-fit tests used in this algorithmic approach, i.e., a stepwise model optimization approach for $k = 30$ items and $n > 300$ respondents with an unknown number of items that might have violated the model’s properties (Debelak, 2018). Accordingly, to reduce the Type-II error of the a posteriori software-based item diagnostic, we chose a Type-I error of 10% ($\alpha = 0.1$) instead of $\alpha = 0.05$ for all tests. Finally, in light of these considerations of statistical power, we also conducted several non-parametric, so-called “quasi-exact” tests (Ponocny, 2001) to verify the prior item selection based on parametric test procedures (Koller et al., 2015). In detail, we selected Ponocny’s (2001) procedures $T_{11}$ as a global test for local dependence, $T_{nd}$ for identifying multidimensionality, $T_1$ for evaluating local dependence between items due to increased inter-item correlations, $T_{in}$ and $T_{it}$ to test for items’ homogeneity and unidimensionality, and $T_{10}$ as a global test of subgroup invariance. Finally, $T_{phis}$ was employed as a test for item discrimination, testing the pointbiserial correlation of each item with the remaining reduced set of items on the scale. For these non-parametric test procedures implemented in the package eRm (Mair et al., 2020), we used $k = 3,000$ sampled matrices for the “burn-in” phase, $m = 1,000$ as the number of excluded matrices in the sampling procedure, and $q = 8,000$ to set the number of effective simulated matrices.

With the exception of one test, indicating that item no. 3 (I 03) showed a significantly lower correlation compared to four of the remaining five items ($p_{T_{phi, 0.05}} = 0.015$), all other non-parametric tests revealed non-significant results. Despite the significant test result ($T_{phi, 03} = T_{phi, I 03}$), we decided to retain item I 03.

Discussion
Our initial study focused on investigating the inner structure of AMMA items in terms of measurement invariance along
Table 3. IRT parameter estimation for AMMA (model no. 1) and model optimization as a function of a stepwise item selection procedure based on goodness-of-fit-test statistics.

| Item | $\beta$ | SE | LL | UL | Step | $\chi^2$ | df | p | Step | z | p | Step | $\chi^2$ | df | p |
|------|--------|----|----|----|------|---------|----|---|------|---|---|------|---------|----|---|
| 101  | 0.83   | 0.11| 0.61| 1.05|      |         |    |   |      |   |   |      |         |    |   |
| 102  | 0.36   | 0.11| 0.14| 0.57|      |         |    |   |      |   |   |      |         |    |   |
| 103  | -2.27  | 0.20| -2.65| -1.89|      |         |    |   |      |   |   |      |         |    |   |
| 104  | 1.26   | 0.12| 1.03| 1.49|      | 8       | 52.18| 21 | <.01|      |   |   |      | 4       | 1.69| 4   |
| 105  | -0.19  | 0.11| -0.41| 0.03|      |         |    |   |      |   |   |      |         |    |   |
| 106  | 0.44   | 0.11| 0.23| 0.66|      | 7       | 64.91| 22 | <.01|      |   |   |      | 4       | 4.19| 4   |
| 107  | 0.42   | 0.11| 0.20| 0.63|      |         |    |   |      |   |   |      |         |    |   |
| 108  | -0.35  | 0.12| -0.58| -0.13|      |         |    |   |      |   |   |      |         |    |   |
| 109  | -0.41  | 0.12| -0.63| -0.18|      |         |    |   |      |   |   |      |         |    |   |
| 110  | -2.40  | 0.21| -2.80| -1.99|      |         |    |   |      |   |   |      |         |    |   |
| 111  | -1.59  | 0.15| -1.89| -1.29|      |         |    |   |      |   |   |      |         |    |   |
| 112  | -0.95  | 0.13| -1.20| -0.70|      |         |    |   |      |   |   |      |         |    |   |
| 113  | -0.41  | 0.12| -0.63| -0.18|      |         |    |   |      |   |   |      |         |    |   |
| 114  | 0.47   | 0.11| 0.25| 0.68| 2     | 115.07| 27 | <.01|      |   |   |      | 2       | 4.55| <.01|
| 115  | -0.80  | 0.12| -1.05| -0.56|      |         |    |   |      |   |   |      |         |    |   |
| 116  | -1.52  | 0.15| -1.82| -1.23|      |         |    |   |      |   |   |      |         |    |   |
| 117  | -1.80  | 0.17| -2.13| -1.48|      |         |    |   |      |   |   |      |         |    |   |
| 118  | 1.50   | 0.12| 1.26| 1.74| 1     | 138.04| 28 | <.01|      |   |   |      | 1       | 5.59| <.01|
| 119  | 0.26   | 0.11| 0.05| 0.48|      |         |    |   |      |   |   |      |         |    |   |
| 120  | -2.16  | 0.19| -2.52| -1.79|      |         |    |   |      |   |   |      |         |    |   |
| 121  | -1.62  | 0.16| -1.92| -1.31|      |         |    |   |      |   |   |      |         |    |   |
| 122  | 1.09   | 0.12| 0.86| 1.31| 3     | 81.23 | 26 | <.01|      |   |   |      | 3       | 3.37| <.01|
| 123  | 0.78   | 0.11| 0.56| 0.99| 14    | 22.53 | 15 | <.10|      |   |   |      | 14      | 2.42| <.02|
| 124  | 1.79   | 0.13| 1.53| 2.04| 12    | 26.77 | 17 | <.06|      |   |   |      | 12      | 2.45| <.01|
| 125  | 1.53   | 0.12| 1.29| 1.77| 11    | 37.13 | 18 | <.01|      |   |   |      | 11      | 3.05| <.01|
| 126  | 1.67   | 0.13| 1.42| 1.92| 6     | 81.80 | 23 | <.01|      |   |   |      | 6       | 4.75| <.01|
| 127  | 0.12   | 0.11| -0.09| 0.34| 10    | 39.99 | 19 | <.01|      |   |   |      | 10      | 2.64| <.01|
| 128  | 1.99   | 0.14| 1.72| 2.26| 9     | 45.27 | 20 | <.01|      |   |   |      | 9       | 2.70| <.01|
| 129  | 0.41   | 0.11| 0.19| 0.62| 4     | 72.48 | 25 | <.01|      |   |   |      | 4       | 3.37| <.01|
| 130  | 1.56   | 0.13| 1.32| 1.80| 5     | 89.81 | 24 | <.01|      |   |   |      | 5       | 4.42| <.01|

Note. a = item excluded from the LR and/or Wald test analysis due to inappropriate response patterns within subgroups. All model analyses as well as model improvements by removing inappropriate items causing a significant model misfit were performed with the R package (Version 1.0-1; Mair et al., 2020) within the statistical environment R (Version 4.0.2 [2020-06-22]; R Core Team, 2020). The median of sample’s raw scores was used as internal split criterion as well as $\alpha = .10$ for all goodness-of-fit tests; global LR-Test revealed a significant deviation between estimated and observed parameter estimations ($\chi^2(28) = 138.04$, $p < .01$), whereas the Martin-Löf test revealed a non-significant result suggesting item homogeneity ($\chi^2(224) = 231.52$, $p = .35$). Items that meet the criteria of all model-fit-tests are highlighted in boldface: $L_{\text{cond}}$ ($n_{\text{par}} = 29$) = -4861.47; $\text{AIC} = 9780.93$, $\text{BIC} = 9893.95$, $\text{cAIC} = 9922.95$.

with homogeneity, unidimensionality, specific objectivity, and their relation to one latent variable by means of the Rasch model. As a main result, the majority of the 30 items showed insufficient psychometric characteristics in accordance with Item Response Theory (Bond et al., 2021). However, we were able to develop a revised 5-item version of AMMA, showing sufficient empirically determined measurement invariance as well as an acceptable reliability—considering the low number of items ($\alpha = 0.637$, 95% CI [0.576, 0.691]). Moreover, despite the exclusion of just over 80% of the items from the original version of AMMA, the internal reliability of the revised scale showed a reliability similar to that of the original subscales ($0.33 \leq \alpha \leq 0.63$, see also Table S5). At this point the question still remained whether these five items included in the final revised test model were sensitive solely to the measurement of audition (and not also to tonal memory; see Introduction). This question was the subject of Study 2.

**Study 2**

**Aims and Hypotheses**

The purpose of the second study was to investigate whether the overall score of the revised version of AMMA exclusively reflected participants’ audition abilities (Gordon, 1989, p. 16) but not their capacity of memorizing music.
Table 4. Parameter estimation for the revised IRT model of the AMMA (model no. 2) after removing items causing a significant model misfit.

| Revised test model | Item fit statistic |
|-------------------|-------------------|
| Item   | β    | SE   | LL   | UL   | $\chi^2$ | df | p    | Outfit MSQ | Infit MSQ | Outfit t | Infit t |
| 103    | -0.46 | 0.20 | -0.85 | -0.08 | 124.957 | 107 | .11 | 1.16 | 1.14 | 0.97 | 1.11 |
| 10     | -0.62 | 0.21 | -1.02 | -0.22 | 85.32 | 107 | .94 | 0.79 | 0.88 | -1.19 | -0.90 |
| 11     | 0.42  | 0.17 | 0.09  | 0.75  | 108.45 | 107 | .44 | 1.00 | 1.01 | 0.08 | 0.19 |
| 16     | 0.52  | 0.17 | 0.19  | 0.85  | 96.94 | 107 | .75 | 0.90 | 0.92 | -1.18 | -1.09 |
| 17     | 0.14  | 0.18 | -0.21 | 0.48  | 105.17 | 107 | .53 | 0.97 | 0.98 | -0.21 | -0.25 |

Note. Stepwise item selection procedure indicated no item for exclusion due to model misfit, that is, all remaining items of the revised model fit the Rasch model. The global LR-Test (split criterion: raw score) revealed a non-significant test result ($\chi^2(3) = 2.18$, $p = .54$); all values of outfit t and outfit t lie within the range from $-2$ to $+2$, indicating an expected fit with the Rasch model (Bond et al., 2021, p. 242); $\Lambda_{\text{GWR}}(\text{fit} = 4) = -193.79$, $\Lambda_{\text{C}} = 395.58$; BIC = 411.17; $\Delta$AIC = 415.17.

(without the need of audiation) as a result of tonal memory as we suspected. Although the item construction of AMMA tends to show a high correspondence with the tonal memory subscale of Seashore’s Measures of Musical Talents (Seashore, 1919), these two measures of musical ability differ concerning their construction of melodic item pairs.

To the best of our knowledge, it is unclear whether both measurement scores (a) show a statistical association and thus a low discriminant validity due to the marked similarity of their item construction, which would point to at least partly identical cognitive processes without a clear distinction between audiation and (short-term) memorization resulting in a positive correlation ($H_1$: $\rho > 0$), or (b) do not show a substantial association – as hypothesized by Gordon (1989) – because of the optimized (i.e., large enough) four-second period of silence between the two melodies of each melody pair ($H_0$: $\rho = 0$).

Consequently, our aims in this study were to compare the discriminant validity of AMMA against tonal memory scores (as measured by the SMMT test). Our study therefore followed a correlational design investigating the strength of the discriminant validity of the revised AMMA version compared to a short version of Seashore’s (1919) tonal memory test.

Method

Sample. In an a priori power analysis ($\rho > 0.30$, $\alpha = 0.05$, $1 - \beta = 80\%$, one-tailed) by means of the software G*Power (Version 3.1.9.3; Faul et al., 2009; Faul et al., 2007), a sample size of $N = 64$ participants was determined as the ideal group size to reveal significant correlations of a moderate effect size according to Cohen’s (1988) conventions. A total of $N = 87$ participants from a German high school ($n = 48$ female [55.2%], $k = 5$ missing) with a mean age of 16.17 years ($SD = 2.53$) ultimately participated in our study. Compared to the norm group of AMMA (Gordon, 1989), our sample showed a slightly lower mean age. However, the age was sufficiently high that all participants were expected to have achieved stabilized audiation abilities (Gordon, 2012). Participants reported playing two instruments on average and having practiced their primary instrument regularly for one hour per day for four to five years. Moreover, our sample showed an average-strength general musical sophistication ($M = 74.33$, $SD = 17.35$, Min. = 36.00, Max. = 111.00) as measured by the German version of the Goldsmiths Musical Sophistication Index, or Gold-MSI (Schaal et al., 2014).

Material. Two measurements were used, one for quantifying participants’ audiation and one measurement for tonal memory. To measure audiation we used the revised 5-item version of the AMMA from Study 1. Although the short scale was Rasch-optimized, we found a somewhat low internal consistency (Cronbach’s $\alpha = 0.44$, see Table S8 in the Supplementary Material section). This can be attributed not only to the items’ quality but also to their small number, the sample size and homogeneity, and the items’ underlying eigenvalue (Bujang et al., 2018; Cortina, 1993; Green et al., 1977; Yurdugül, 2008). As indicated by the scale mean ($M = 0.88$, $SD = 0.18$), all five items tended to be insufficiently difficult for our sample ($0.807 \leq M \leq 0.964$), along with item-total correlations that were in an almost acceptable range ($0.211 \leq r \leq 0.276$).

Given the time constraints for the entire test procedure in the schools, we selected five items from the tonal memory subscale of the German SMMT (Seashore, 1919) to create a short version of the original test. The item selection was based on the difficulties reported in the manual for the German version (Butsch & Fischer, 1966) and resulted in a uniform distribution of item difficulties. For the short version of SMMT we selected items A6, B2, B8, C4, and C10. Post hoc descriptive item and reliability analyses of the SMMT short scale for the measurement of tonal memory (Table S8 in the Supplementary Material section) revealed that items were generally of a medium to low difficulty level ($0.736 \leq M \leq 0.920$). Compared to our
revised AMMA version, the short version of the SMMT tonal memory scale showed a stronger and acceptable internal consistency (Cronbach’s $\alpha = 0.63$) considering the small number of items and their item-total correlations ($0.228 \leq r \leq 0.468$).

All melody pairs were presented by means of compact disc. A silent period of 3.3 s was set as a response time for the short form of the tonal memory test, after which the next melody pair was announced and presented immediately. For the revised AMMA version, participants gave immediate responses within a silent period of 2 s inserted after each melody pair. Three practice examples were presented for each of the measurements. While the original practice examples were used for the short version of AMMA, the three items A5, B5, and C5 from SMMT were adopted as practice items for measuring tonal memory.

**Procedure.** Data were collected in class sizes of $n \leq 30$ participants in a secondary school in Germany. Data collection was similar to the procedure of Study 1. However, due to time constraints, each session was limited to 20 min.

First, participants were welcomed and informed about the objectives of the study. After giving written informed consent, participants carefully read the instructions of the short version of the AMMA. Next, all three practice examples were presented, followed by the evaluation of the five test items of the revised AMMA scale from Study 1. Afterwards, participants read the instructions of the subsequent SMMT test for tonal memory before listening to the practice examples, followed by the presentation of five items of the SMMT short scale. Finally, participants gave a self-disclosure about their musical background based on the general factor items of the Goldsmiths Musical Sophistication Index (Müllensiefen et al., 2014). No reimbursement was paid. Participants were informed about their test performance two weeks later.

**Data Analysis.** The data analysis followed a two-step approach. First, participants’ responses given on all items of both measurements (AMMA and SMMT) were dichotomized ($0 =$ incorrect answer; $1 =$ correct solution) following Gordon’s procedure for investigating construct validity (Gordon, 1991, p. 8). Next, using JASP (Version 0.13.1, JASP Team, 2020), we conducted item and scale reliability analyses for both scales (see Table S8 in the Supplementary Material section) and calculated overall scores of both measurements by summing up all the correct answers. Due to missing responses for one or more items of AMMA, the response sheets of four participants were excluded from further data analyses. Based on the remaining 83 response sheets, a correlation between total scores of both measurements was calculated (“raw correlation”). To arrive at the latent “true” correlation as a quantification of the statistical association between both variables if both measurements had shown perfect reliability, raw correlations were corrected for attenuation effects (Hunter & Schmidt, 2004). Because of a ceiling effect of the overall scores of both measurements we used a non-parametric correlation analysis. Finally, in addition to the non-Bayesian frequentist correlation approach, we conducted a Bayesian post hoc analysis using JASP (Version 0.13.1, JASP Team, 2020) with default settings for less informative priors on the correlation corrected for artifacts (“summary statistics”).

**Results**

Based on the 83 valid cases, we found a significant positive raw correlation between SMMT tonal memory measurements and AMMA audiation scores of moderate effect size ($r = 0.30$ [0.22; 1.00], $p < 0.001$, one-tailed). Taking into account the reliability of each test (which had an effect on the reduction of the observed “raw” correlation), the “true” correlation corrected for artifacts revealed a stronger association between both test scores ($r = 0.58$, $p < 0.001$).

The post hoc Bayesian analysis provided further insights. The Bayesian parameter estimation of the disattenuated correlation between audiation and tonal memory showed a small 95%-credible interval in addition to the point estimator (Mdn$_{r} = 0.559$, 95% CI [0.418; 0.70]). In other words, given our data, there was a 95% probability that the population correlation between both test results would fall within the credible interval [0.418; 0.70], thus showing a moderate to strong effect size (for details, see Figure 1). Since the 95%-credible interval did not include $\tau = 0$, there was practically no evidence for a null correlation. This conclusion was also supported by the Bayes factor indicating that there was “decisive” evidence according to Jeffreys’ benchmarks for the strength of evidence (Jeffreys, 1961) in favor of the alternative hypothesis ($BF_{10} = 2.67 \cdot 10^{12}$) compared to the null hypothesis. Thus, our alternative hypothesis was $2.67 \cdot 10^{12}$ times more likely compared to the null hypothesis which Gordon (1989) proposed.

![Figure 1. Prior and posterior probability as well as the Bayes factor of the disattenuated correlation between overall scores for the AMMA and SMMT tonal memory subtest for the alternative hypothesis (JASP team, 2020).](image)

$BF_{10} = 2.67e+12$
$BF_{01} = 3.75e-13$

Median: 0.559
95% CI: [0.418, 0.70]
A Bayes factor robustness check indicated that the choice of the default less-informative prior ($\kappa = 1$) did not yield the maximum Bayes factor, which would have only been achieved by $\kappa = 1.493$ ($BF_{+0} = 1.829 \cdot 10^{17}$). Moreover, as can be seen from Figure 2, even if the shaping of the a priori probability had suggested strong evidence for the discriminant validity of AMMA along with a narrow 95%-credible interval by choosing $0.003 \leq \kappa \leq 0.1$, the Bayes factor would still have been in a range between 41.35 $\leq BF_{+0} \leq 1.85 \cdot 10^{10}$. This, according to Jeffreys’ (1961) suggested interpretation of the Bayes factor, would still have represented “very strong” to “decisive” strength of evidence for the alternative hypothesis. Therefore, it is reasonable to assume that no bias in favor of the alternative hypothesis occurred as a result of our opting for less informative priors.

**General Discussion**

There were two main aims of this investigation: first, we wanted to evaluate the item characteristics of AMMA in terms of up-to-date psychometric standards so we could draw conclusions about the quality of the AMMA test (Study 1). Second, we wanted to evaluate AMMA’s construct validity to test Gordon’s claim that AMMA is an inventory for the exclusive measurement of audiation as a latent variable (Study 2). If this latter assumption were true, only a (very) small correlation should have been observed between AMMA scores and scores from analogously constructed tests for tonal memory (in the present study the tonal memory subtest of the SMTT).

In Study 1, we focused on participants’ overall test scores. Although our sample performed similarly to that of the norm group reported by Gordon (1989), our results revealed a slightly lower internal test reliability for the total test score ($\alpha_{TS} = 0.72$, see Table S4) compared to recommended reliability standards for state-of-the-art psychometric measurements (Abell et al., 2009, pp. 97–95). Moreover, the majority of the subscales’ internal consistencies was much lower than that of the overall test score ($\alpha_{same} = 0.33; \alpha_{rhythm} = 0.48; \alpha_{tonal} = 0.60$, see Table S5). Because of these insufficient reliability coefficients, the AMMA scores showed a large standard error of measurement ($SEM = 2.28; \alpha_{TS} = 0.72; SD_{TS} = 4.30$), significantly exceeding the “5% or less rule” of the overall test scores’ range as proposed by Abell et al. (2009, p. 96).

As a consequence of using an algorithm-based model optimization strategy, $k = 25$ items had to be excluded in order to obtain a Rasch model with a non-significant model misfit. Along with the exclusion of items, a significant decrease in the model information criteria was observed, indicating an improvement in the quality of the final as compared to the initial model (see Tables 3 and 4). Following from that, the revised version of AMMA (model No. 2) contained the five remaining items of which three melody pairs had rhythmic variations (I 03, I 10, and I 17) and two pairs featured tonal variations (I 11 and I 16). Interestingly, the final model did not include any “same” items with no difference between the two melodies.

Although this optimization followed an objective and statistically driven approach for the enhancement of AMMA’s measurement invariance as part of meeting construct validity, the probability of ceiling and floor effects increased significantly due to the small number of remaining items. Moreover, the risk of ceiling and floor effects was compounded by the fact that the variance in the difficulty of the remaining items was low and fell within a narrow range of a medium skill/task level demand (Table 4).

Another consequence was that AMMA in its revised version would probably not be a suitable tool for identifying people with high audiation abilities as an indicator of “giftedness”. In other words, the difficulty of the remaining five test items was too low with ensuing ceiling effects. For example, someone with an average level of audiation ($\theta = 0$) would have a probability of 37.3% for giving a correct response on the remaining item with the highest difficulty ($J_{116} = 0.52, 95\% CI [0.19, 0.85]$), or a person with an audiation ability of only one standard deviation above the population average would have a 61.8% probability of giving a correct response on the same most difficult item. To discriminate between a participant with a significantly superior level of audiation (e.g., $\theta \geq 2$) and another with an above-average level of audiation (i.e., $1 \leq \theta < 2$), the revised model should ideally include items with difficulty values corresponding to the full range of participants’ abilities. However, if used to identify people with a music ability far above average, AMMA ought to include more items of higher difficulty ($\sigma_{i} > 1$) and fewer of lower difficulty. Interestingly, even in the original version of the AMMA, only eight out of 30 items (26.7%) showed a difficulty level of $\sigma_{i} > 1$ (see Table 3), a proportion which is rather low for a measurement intended to differentiate between
groups of various ability levels (Gordon, 1989, p. 16), including people with stabilized audiation far above the population average. On the contrary, a higher proportion of items with a medium to high difficulty level would have been expected if the measurement is supposed to (also) differentiate between groups of average music ability and groups showing greater music ability (in terms of “giftedness”).

The large number of excluded items due to model misfit might be plausible in light of earlier studies focusing on the latent-factor structure of the AMMA test. For example, in 1989 Gordon expected one general factor but was unable two years later to identify it. Instead, he found nine factors accounting for no more than 45% of variance. His own results in essence support our assumption that AMMA’s items might be characterized by great heterogeneity, leading to a multidimensionality rather than a unidimensionality in test scores. Moreover, as in our study, the items on the strongest factor accounted for a low proportion of variance (13.4% of the variance, and only 11.2% after Varimax rotation; Gordon, 1991, p. 17). Likewise, Gordon (1991, pp. 1–21) and Verdis and Sotiriou (2018) were unable to identify a one-dimensional solution for the latent structure underlying AMMA, despite their using different statistical approaches. In essence, the large number of excluded items does not come as a surprise, but rather confirms earlier findings. But even with few items the revised scale does not lose much reliability (Cronbach’s α = 0.63) compared to the reliabilities of the original AMMA subscales. We view this as a sign of a successful revision of the test model in our first study.

In Study 2, the main result of a high correlation between the score from the revised AMMA version and five selected items from the tonal memory subtest of the SMMT stood in contrast to Gordon’s (1989) assumption that effects based on short-term memory would not play any role in the measurement of audiation as long as a four second long gap of silence separated the two melodies in a melody pair. Yet all reliability of audiation as long as a four second long gap of short-term memory would not play any role in the measurement scores in Study 2 might reflect how audiation depends on the ability of memorizing music as a function of (short-term) memory, which is especially characteristic for the early stages of audiation. One could speculate that more advanced types of audiation would yield lower correlations with memory. But some reliance on tonal memory seems inevitable.

Moreover, due to the AMMA test design, participants only have to discriminate between the two melodies of a melody pair to give a correct response on the items. They do not require further domain- or task-specific knowledge. This results in a perceptually and not audiation-driven “same or different” decision (with respect to tonal or rhythmic structure). From a theoretical perspective, however, there are two reasons why this kind of perceptually driven discrimination task has a reduced diagnostic value for measuring music ability: (1) Solving basic auditory discrimination tasks without further need for content- or task-knowledge is not restricted to human behavior but also occurs in animals (e.g., Brooks & Cook, 2010; Cook et al., 2016; Hagmann & Cook, 2010; Porter & Neuringer, 1984; Watanabe & Sato, 1999). Thus, the item design refers to a cognitive skill that is not unique to human behavior but which is also manifested in certain animal species. Nonetheless, such a paradigm can be applied convincingly when the content of the items mirrors the cultural specifics of the particular skill and the respective test items are so distinctive that they require the use of the relevant domain-specific skill or, inversely stated, if basic perceptual processing is not sufficient to solve the items. (2) Perceptually driven aural discrimination is a general cognitive skill which is not domain-specific. For example, in a study imposing similar demands on the participants as in our studies, Stepanov et al. (2020) demonstrated that auditory discrimination is a general cognitive process primarily determined by the capacity of working memory. They asked 70 children (mean age 9.17 years) in a prosodic discrimination task to decide whether there was a difference within pairs of sentences (K = 70) presented in French, a language the children neither spoke nor understood. Here, children’s discrimination accuracy depended only on their success in extracting and processing acoustic features – and not on detecting semantic differences. Children’s discrimination accuracy correlated strongly with their performance on working memory measures such as forward digit span tasks (R = 0.78, p = 0.0026), backward digit span tasks (R = 0.92, p = 0.0013), and non-word repetition tasks (R = 0.39, p = 0.017). Our correlation between audiation and tonal memory was of similar magnitude (MDn = 0.559, 95% CI [0.418; 0.701]) and might be explained by participants’ working memory capacity which could be responsible for their time- and pitch-related
auditory discrimination accuracy in both measurements. As stated above, it is likely that participants are able to solve many of the AMMA’s items without the domain-specific skill of audiation and are instead merely using basic auditory processing skills, including working memory. This question could be investigated in future studies that control for general working memory capacity.

Although we found convincing evidence for deficient psychometric properties of AMMA, such as low internal validity, low reliability, heterogeneous item characteristics, underlying multi-dimensional latent structure, and low construct validity, we cannot pass final judgment on the existence of audiation as a latent variable. However, compared to previous critical evaluations of Gordon’s test development, we do not think that AMMA’s fundamental problem of lacking validity can be resolved by re-norming the test (as suggested by Grashel, 2008). We are not convinced that observed weaknesses in test construction should be attributed to differences in societal contexts when the test was designed, nor to “characteristics of the music students in the 21st century” (Hanson, 2019, p. 208). Instead, we argue for a more radical renewal. We are convinced that, in its current form, AMMA is unsuitable for a valid and reliable operationalization of the audiation construct. Measuring audiation thus calls for new diagnostic instruments to be developed that focus on the multi-faceted structure of audiation with more consideration of domain-specific tasks and the cognitive processes that underlie them. In recent years, several tools have been developed for measuring skills related to the types and stages of the phenomenon of audiation, such as notation-evoked sound imagery (e.g., Wolf et al., 2018) or auditory mental imagery (e.g., Gelding et al., 2020). It is time to develop a new series of tests for measuring audiation that adhere to the high standards of today’s psychometric testing, including Rasch model conformity, automatic item generation (Gierl & Haladya, 2013), and adaptive testing (Harrison et al., 2017).

Acknowledgements
The authors would like to thank Maria Lehmann for the manuscript editing; Marcus Büring for his support in data collection; Johannes Hasselhorn and Fanny Empacher for helpful comments on earlier versions on the manuscript.

Contributorship
FP and RK conceived the study. All authors were involved in study design, data collection, and analysis. FP wrote the first draft of the manuscript. All authors reviewed and edited the manuscript and approved the final version of the manuscript.

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Action editor
Graham Welch, University College London, Institute of Education.

Peer review
Adam Ockelford, University of Roehampton, Applied Music Research Centre.
Peter Webster, University of Southern California, School of Music.

ORCID iDs
Friedrich Platz https://orcid.org/0000-0002-9758-9434
Reinhard Kopiez https://orcid.org/0000-0003-3356-3478
Anna Wolf https://orcid.org/0000-0002-6196-6488

Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material
Supplemental material for this article is available online. Additional data resources are available at https://osf.io/gz4sv.

References
Abell, N., Springer, D. W., & Kamata, A. (2009). Developing and validating rapid assessment instruments. Oxford University Press.
American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). Standards for educational and psychological testing. American Educational Research Association.
Andersen, E. B. (1973). A goodness of fit test for the Rasch model. Psychometrika, 38(1), 123–140. doi:10.1007/BF02291180
Bond, T. G., Yan, Z., & Heene, M. (2021). Applying the Rasch model: Fundamental measurement in the human sciences (4th ed.). Routledge.
Brooks, D. I., & Cook, R. G. (2010). Chord discrimination by pigeons. Music Perception, 27(3), 183–196. doi:10.1525/mp.2010.27.3.183
Bugos, J. A., Perlstein, W. M., McCrae, C. S., Brophy, T. S., & Bedenbaugh, P. H. (2007). Individualized piano instruction enhances executive functioning and working memory in older adults. Aging & Mental Health, 11(4), 464–471. doi:10.1080/13607860601086504
Bujang, M. A., Omar, E. D., & Baharum, N. A. (2018). A review on sample size determination for Cronbach’s Alpha test: A simple guide for researchers. Malaysian Journal of Medical Sciences, 25(6), 85–99. doi:10.21315/mjms2018.25.6.9
Burgoine, A. P., Harris, L. J., & Hambrick, D. Z. (2019). Predicting piano skill acquisition in beginners: The role of general intelligence, music aptitude, and mindset. Intelligence, 76, 101383. doi:10.1016/j.intell.2019.101383
Butsch, C., & Fischer, H. (Eds.). (1966). Seashore-Test für musikalische Begabung. Testanweisung [German manual of the Seashore measures of musical talent]. Verlag Hans Huber.
Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Psychology Press.
Colwell, R. (2002). Assessment’s potential in music education. In R. Colwell & C. Richardson (Eds.), The new handbook of
Jeffreys, H. (1961). Theory of probability (3rd ed.). Oxford University Press.

Koller, I., Alexandrowicz, R., & Hatzinger, R. (2012). Das Rasch-modell in der praxis. Eine einführung mit eRm [The Rasch model in practice. An introduction with eRm]. Facultas Verlags- und Buchhandels AG.

Koller, I., Maier, M. J., & Hatzinger, R. (2015). An empirical power analysis of quasi-exact tests for the Rasch model. Methodology, 11(2), 45–54. doi:10.1027/1614-2241/a000090

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

Mair, P., Hatzinger, R., & Maier, M. J. (2020). eRm: Extended Rasch Modeling (Version 1.0-1) [Computer Program]. https://cran.r-project.org/package=eRm

Martin-Löf, P. (1973). Statistika modeller [Statistical models. Notes from seminars 1969-70 by R. Sundberg, 2nd. ed.]. Facultas Verlags- und Buchhandels AG.

Müllensiefen, D., Gingras, B., Musil, J., Stewart, L., & Snyder, J (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

Ollen, J. E. (2006). A criterion-related validity test of selected indicators of musical sophistication using expert ratings [Dissertation]. Ohio State University. Electronic Theses & Dissertations Center. http://rave.ohiolink.edu/etdc/view?acc_num=osu1161705351

Ponocny, I. (2001). Nonparametric goodness-of-fit tests for the Rasch model. Psychometrika, 66(3), 437–459. doi:10.1007/BF02294444

Porter, D., & Neuringer, A. (1984). Music discriminations by pigeons. Journal of Experimental Psychology: Animal Behavior Processes, 10(2), 138–148. doi:10.1037/0097-7403.10.2.138

R Core Team. (2020). R: A language and environment for statistical computing (Version 4.0.2 – “Taking Off Again”) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Sang, R. C. (1998). A four-year study of the predictive validity of the advanced measures of music audiation. Southeastern Journal of Music Education, 10, 127–138.

Schaal, N. K., Bauer, A.-K. R., & Müllensiefen, D. (2014). Der gold-MSI: Replikation und validierung eines fragebogeninstruments zur messung musikalischer erfahrenheit anhand einer deutschen stichprobe [The gold-MSI: replication and validation of an assessment for the measurement of musical sophistication in a German sample]. Musicae Scientiae, 18(4), 423–447. doi:10.1177/1029864914541851

Schellenberg, E. G., & Weiss, M. W. (2013). Music and cognitive abilities. In D. Deutsch (Ed.), The psychology of music (3rd ed., pp. 499–550). Academic Press.

Schleuter, S. L. (1993). The relationship of AMMA scores to sight singing, dictation, and SAT scores of university music majors. Contributions to Music Education, 20, 57–63. https://www.jstor.org/stable/24127331

Schneider, P., Scherg, M., Dosch, H., Specht, H. J., Gutschalk, A., & Rupp, A. (2002). Morphology of Heschl’s gyrus reflects enhanced activation in the auditory cortex of musicians. Nature Neuroscience, 5(7), 688–694. doi:10.1038/nn871

Schneider, P., Sluming, V., Roberts, N., Bleeck, S., & Rupp, A. (2005). Structural, functional, and perceptual differences in Heschl’s gyrus and musical instrument preference. Annals of the New York Academy of Sciences, 1060, 387–394. doi:10.1196/annals.1360.033

Seashore, C. E. (1919). The psychology of musical talent. Silver, Burdett and Company.

Shrout, P. E., & Lane, S. P. (2012). Reliability. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), APA Handbook of research methods in psychology (Vol 1, pp. 643–660). American Psychological Association.

Shuter-Dyson, R. (1999). Musical ability. In D. Deutsch (Ed.), The psychology of music (2nd ed., pp. 627–651). Academic Press.

Stepanov, A., Kodrič, K. B., Stateva, P., & Strijkers, K. (2020). The role of working memory in children’s ability for prosodic discrimination. PLoS One, 15(3), e0229857. doi:10.1371/journal.pone.0229857

Stone, C. A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. Journal of Educational Measurement, 40(4), 331–352. doi:10.1111/j.1745-3984.2003.tb01150.x

Tan, Y. T., McPherson, G. E., Peretz, I., Berkovic, S. F., & Wilson, S. J. (2014). The genetic basis of music ability. Frontiers in Psychology, 5, 658. doi:10.3389/fpsyg.2014.00658

Verdis, A., & Sotiriou, C. (2018). The psychometric characteristics of the advanced measures of music audiation in a region with strong non-western music tradition. International Journal of Music Education, 36(1), 69–84. doi:10.1177/0255761417689925

Vispoel, W. P. (1993). Recent research in computerized adaptive testing of musical aptitude. Update: Applications of Research in Music Education, 11(2), 39–42. doi:10.1177/875512339301100209

Vispoel, W. P., Wang, T., & Bleiler, T. (1997). Computerized adaptive and fixed-item testing of music listening skill: A comparison of efficiency, precision, and concurrent validity. Journal of Educational Measurement, 34(1), 43–63. doi:10.1111/j.1745-3984.1997.tb00506

von Davier, M. (1997). Bootstrapping goodness-of-fit statistics for sparse categorical data: Results of a Monte Carlo study. Methods of Psychological Research, 2(2), 29–48. http://www.dgps.de/fachgruppen/methoden/mpr-online/

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large.
Transactions of the American Mathematical Society, 54(3), 426–482. doi:10.2307/1990256

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

Walters, D. L. (1989). Audiation: The term and the process. In D. L. Walters & C. C. Taggart (Eds.), Readings in music learning theory (pp. 3–11). GIA Publications.

Watanabe, S., & Sato, K. (1999). Discriminative stimulus properties of music in Java sparrows. Behavioural Processes, 47(1), 53–57. doi:10.1016/s0376-6357(99)00049-2

Watson, D. (2012). Objective tests as instruments of psychological theory and research. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), APA Handbook of research methods in psychology (Vol. 1, pp. 349–369). American Psychological Association.

Wolf, A., & Kopiez, R. (2018). Development and validation of the musical ear training assessment (META). Journal of Research in Music Education, 66(1), 53–70. doi:10.1177/0022429418754845

Wolf, A., Kopiez, R., & Platz, F. (2018). Thinking in music: An objective measure of notation-evoked sound imagery in musicians. Psychomusicology: Music, Mind, and Brain, 28(4), 209–221. doi:10.1037/pmu0000225

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. Mesa Press.

Yurdugül, H. (2008). Minimum sample size for Cronbach’s coefficient alpha: A Monte-Carlo study. Hacettepe Üniversitesi Journal of Education, 35, 397–405.