You *Can* Tell a Prodigy From a Professional Musician: A Replication of Comeau et al.'s (2017) Study

Viola Pausch, Nina Düvel, & Reinhard Kopiez
Hanover University of Music, Drama and Media, Hanover, Germany

According to Feldman (1993), musical prodigies are expected to perform at the same high level as professional adult musicians and, therefore, are indistinguishable from adults. This widespread definition was the basis for the study by Comeau et al. (2017), which investigated if participants could determine whether an audio sample was played by a professional pianist or a child prodigy. Our paper is a replication of this previous study under more controlled conditions. Our main findings partly confirmed the previous findings: Comparable to Comeau et al.'s (2017) study (N = 51), the participants in our study (N = 278) were able to discriminate between prodigies and adult professionals by listening to music recordings of the same pieces. The overall discrimination performance was slightly above chance (correct responses: 53.7%; sensitivity d' = 0.20), which was similar to Comeau et al.'s (2017) results of the identification task with prodigies aged between 11 and 14 years (approximately 54.6% correct responses; sensitivity approximately d' = 0.13). Contrary to the original study, musicians and pianists in our study did not perform significantly better than other participants. Nevertheless, it is generally possible for listeners to differentiate prodigies from adult performers—although this is a demanding task.

Received: July 13, 2021, accepted May 28, 2022.

Key words: musical prodigy, signal detection theory, replication study, wunderkind, precocity

Remarkable musical performances of very young children have been the subject of reports and investigations since the Age of Enlightenment. For example, Barrington (1770) studied the 8-year-old Wolfgang Amadeus Mozart during a concert tour in England. Barrington carried out a number of musical performance tests (e.g., sight reading in various clefs and harmonic modulations) and approved little Mozart's extraordinary musical talent. In the early 20th century, musical child prodigies came into the focus of psychology: Richet (1900) published a report on a number of musical tests with the 3-year-old José Rodriguez (Pepito) Arriola (for a comprehensive description, see Graus, 2021), and the German psychologist Baumgarten (1930) developed a standard psychological screening for the objective measurement of skills of prodigious children (from 6 to 14 years of age) from various fields (for an overview of the beginnings of an objective psychological assessment of musical prodigies, see Kopiez & Lehmann, 2016). Musical prodigies were also of interest in empirical musicology: For example, Kopiez (2011) analyzed a sample of 213 Europe-wide reports published between 1798 and 1848 and found an average age of 10.73 years (range: 4–16 years) at the first public performance. The average age of debut was invariant over the observed period of 50 years regardless of the instrument or gender.

However, the question remains whether there is a specific age at which the performance of the child prodigy is comparable to that of an adult professional performer, thus making it impossible to determine if the performer is actually an adult. In their comparative overview of prodigy definitions, Marion-St-Onge et al. (2021, Supplementary Table 1) show that there is not a clear age limit. Based on their review of various sources, the age range of interest must be somewhere between under 10 years and the onset of puberty (for girls, between 10 and 11, and for boys, between 11 and 12 years). Interestingly, in their sample of 19 current and former prodigies, the authors found an average age of 10.3 years (SD = 1.8, range = 7–13) when prodigy status was reached. This finding is very similar to the result of 10.73 years for the age of the first public performance observed in the aforementioned historiometric analysis by Kopiez (2011).

In his long-term study of six remarkable children, Feldman (1986) decided on 10 years of age as the upper age limit for his subjects. This age threshold for a child

Music Perception, volume 40, issue 1, pp. 39–54, issn 0730-7829, electronic issn 1533-8312. DOI: https://doi.org/10.1525/mp.2022.40.1.39

to be classified as a prodigy is not explained further and thus seems to be set arbitrarily and heuristically. Feldman and Morelock (2011, p. 212) only state that this definition "was intended to guide research and, at the same time, to be explicit and precise enough to be tested empirically." It served as a base for further definitions such as those of Solomon ("before the age of twelve," 2012, p. 405) and Shavinina (2016, p. 259).

Against this background, Feldman's (1993, p. 188; 1986, p. 16) prominent definition of a prodigy as being a child younger than 10 years of age who can perform at the same high level as a professional adult in an intellectually demanding field seems to be a reasonable starting point for our investigation. In the case of music, such a child is described as a musical prodigy or wunderkind (borrowed from the German language). Based on Feldman and Morelock's (2011, p. 212) aforementioned idea to guide research by providing a first definition that includes the age limit of 10 years, this age limit was not followed rigorously for the present study due to its arbitrariness. Moreover, the present study is not meant to define what characterizes a musical prodigy but rather focuses on the aspect that performances of child prodigies and adult professionals are assumed indistinguishable. Thus, it is a replication of the study by Comeau et al. (2017), in which the researchers examined whether musical child prodigies performed at the same high level as adults, as outlined in Feldman's above definition. Accordingly, in their online study, 51 participants (26 musicians and 25 nonmusicians) determined whether 165 audio examples were played by prodigies between the age of 7 and 14 or by professional adult pianists. The underlying null hypothesis was that such discrimination would not be possible and that the results would therefore be in line with Feldman's definition. Contrary to this hypothesis, however, the results showed that musicians and nonmusicians were able to distinguish between prodigies and adults above chance "but by a very modest margin" (Comeau et al., 2017, p. 200): Musicians identified the player (prodigy or adult professional) correctly in 62.6% (SD = 7.5) of the recordings, and nonmusicians identified the player correctly in 52.8% (SD = 5.0) of the same recordings (Comeau et al., 2017, p. 202). Overall, this led to a correct response rate of 57.8%. In the original study, Comeau et al. (2017) used the age limit of 10 years for the younger prodigies. They employed musical examples of prodigies between 7 and 10 years of age as well as between 11 and 14 years of age. Regarding only the younger prodigies (7 to 10 years of age)—in other words, those in line with Feldman's definition—the correct response rate increased to approximately 60.5% in Comeau et al.'s study. The tasks comprising older prodigies aged between 11 and 14 years yielded approximately 54.6% correct responses (sensitivity approximately d' = 0.13; musicians: approximately 59.5%, nonmusicians: approximately 49.5%; see last column of Table 6).

Even if the method of comparing child prodigies and adult musicians was carefully designed in Comeau et al.'s (2017) study, some questions arising from the original study remain open and require further consideration. First, lengths of the stimuli varied considerably between 15 and 128 s. Second, the sound quality of some stimuli (e.g., studio recordings) was deliberately degraded by technical means, while for other stimuli of lower quality (e.g., live concert recordings obtained from YouTube), the quality was maintained. Thus, the sound quality of the stimuli was poor, and the length was not consistent. Therefore, it was difficult to compare the stimuli. Moreover, the audio quality could also have a subliminal influence on the decision making, which would lead to noisy data. To clearly answer the research question of the study, professionally recorded stimuli of equal length and quality would have been a better strategy for stimulus selection.

Furthermore, the selection of audio stimuli may have been confounded by specialization and expertization effects. Expertization refers to the phenomenon that the average level of professional musicians has increased over the last centuries. This effect is caused by the following processes: In the history of music performance, performers have often been motivated to achieve increasingly higher and more complex levels of performance as long as there are incentives, such as outperforming their peers or overcoming previous practical–technical limits (Lehmann, 2006, pp. 3–4). Therefore, improved instruments were built (Lehmann, 2006, pp. 7, 13), and musicians began to specialize in one field such as performance or composition (Lehmann, 2006, pp. 7, 19). For a more efficient musical training, which was also accessible to a broader audience, innovative playing techniques were developed over longer periods of time (Lehmann, 2006, p. 13). Nowadays, compared to previous generations, musicians spend more time for the preparation of performances (Lehmann, 2006, p. 19). In turn, concerts take place in front of a more sophisticated and therefore more demanding and critical audience than in the past, which can even compare live performances with recordings (Lehmann, 2006, p. 10). This sets high expectations for the performer and, in the best case, might also increase their level of performance. All these developments, also referred to as long-term expertization, suggest that the skills of musicians—at least at an expert level and in the

domain of popular solo instruments—have steadily improved over the last three centuries (Lehmann, 2006, p. 4). As elaborated by Lehmann (2006, p. 17), this also applies to musical child prodigies. He found that there was a significant tendency for prodigies from later generations to play more difficult pieces after a shorter period of practice (hence, at a younger age) compared to child prodigies from previous generations ($r = .67$, $p < .05$). Additionally, child prodigies from later generations were more advanced in their abilities than those from earlier generations ($r = .85$, $p < .01$; Lehmann, 2006, p. 17). In summary, the author concludes that prodigies of our time play more complex pieces and that precocity has increased over time (Lehmann, 2006, p. 17). This would mean that today's child prodigies are likely to have a higher level of expertise compared to child prodigies in former times. Therefore, one should always consider a wunderkind in relation to the standards of their historical time (Olbertz, 2010, p. 529). In so doing, one avoids the irresolvable question whether potentially similarly high performances of prodigies and adult musicians of different generations are caused by the above-mentioned specialization and expertization effect or by the fact that prodigies really perform at the same high level as even contemporaneous adult professional pianists of their generations. In other words, today's child prodigies should not be compared with well-known pianists of earlier days. However, this is exactly what Comeau et al. (2017) did: Some of the first-class pianists from the study had been born as early as in the 19th century or at the beginning of the 20th century; e.g., Artur Schnabel (1882–1951), Wilhelm Backhaus (1884–1969), Wilhelm Kempff (1895–1991), Walter Gieseking (1895–1956), Jacques Février (1900–1979), Vladimir Horowitz (1903–1989), Claudio Arrau (1903–1991), Emil Gilels (1916–1985), and Sviatoslav Richter (1917–1997). In contrast, we decided to select only professional pianists who are currently still performing.

STUDY AIMS

From our perspective, these methodological shortcomings and possible confounding effects caused by different recording qualities (which might have inflated the Type I error), generations of players (which might have inflated the Type II error), and a very long testing procedure for each participant (which might have inflated the Type II error) justified a replication of Comeau et al.'s (2017) study. Using Feldman's definition of a child prodigy, we went from the null hypothesis that it would not be possible to decide above chance whether an audio sample was played by a professional adult or a child prodigy. We wanted to test whether Comeau et al.'s main result of the discriminability of child prodigies and adults could be replicated or whether it was an artefact caused by the different confounding factors. In the end, this would have consequences for the definition of a musical prodigy, especially the aspect of an age threshold. An unambiguous definition would be a contribution to theory formation and of practical importance for future research.

HYPOTHESES

The following empirical hypotheses were investigated:

A. Overall, participants can discriminate between child prodigies and adult professionals.
B. Professional and amateur musicians can better discriminate between prodigies and adult professionals than nonmusicians.
C. Professional pianists show the highest discrimination performance.
D. There is a correlation between discrimination ability and the degree of musical sophistication measured by the Goldsmiths Musical Sophistication Index (Gold-MSI; Müllensiefen et al., 2014; Schaal et al., 2014).

Hypotheses A and B were also investigated by Comeau et al. (2017). Hypotheses C and D were added so that we could gain deeper insights into a possible influence of expertise on classification performance.

Method

PARTICIPANTS

Professors and students of music departments at German universities of music were invited by mailing lists to take part in the online survey. Additionally, the survey was open to other interested people, who were invited via email, mailing lists, and social media (e.g., piano, music students, or musician groups on Facebook).

In the end, 568 people started the online questionnaire, 275 (48.4% of those who started) completed the entire questionnaire, and 279 completed all tests and questions except for the Headphones and Loudspeaker Test (HALT; Wycisk et al., 2022; Wycisk et al., 2018). As recommended by Leiner (2019), participants with a high relative speed index (1.75 or greater) were excluded. That was the case for one person. Finally, valid data from $N = 278$ participants were obtained for the main analysis, which included all tests and questions except for HALT (162 women, 108 men, 8 diverse or not specified; age: $M = 31.4$ years, $SD = 12.1$, range = [17, 73]). To increase the research range, the questionnaire was

translated into Traditional Chinese. The decision for Traditional Chinese was motivated by the availability of the Chinese version of the Gold-MSI and because one of our cooperation associates being a native Traditional Chinese speaker. This enabled us to reach another audience from a more global perspective (including participants mainly from Taiwan). The Traditional Chinese version was answered by 110 (39.6%) participants.

Participants were informed that taking part in the study was voluntary. They gave informed consent that the collected data could be used for scientific purposes as well as for storage and analysis in anonymized form.

Statistical Power and Required Sample Size

Calculations of the required sample size (see Section A in the Supplementary Material accompanying the online version of this paper at mp.ucpress.edu for details) showed that it would have been sufficient to collect data from 198 participants to obtain a power of at least $1 - \beta = .95$ for McNemar's test (McNemar, 1947). However, due to a higher number of actual participants, the test power was $1 - \beta = .989$ for Miettinen's (1968) approach and $1 - \beta = .988$ for Bennett and Underwood's (1970) approach. Consequently, sample size and, therefore, also test power of the study exceeded the minimum requirements according to the standards of $\alpha = .05$ and $1 - \beta = .80$ (conventional benchmarks as recommended by Cohen, 1988, p. 56; Ellis, 2010, p. 53).

Expertise

Participants were asked to declare whether and at what level they played the piano, and to self-assess their level of musicianship on a 3-point scale: 1) nonmusician, 2) amateur musician, or 3) professional musician (for the exact wording, see Table D2 in the online Supplementary Material at mp.ucpress.edu). The obtained sample showed high piano affinity: 68.3% played the piano at an amateur or professional level. From the total of $N = 278$ participants, 11.9% were self-declared professional pianists, and 10.8% were professional musicians with other main instruments. The largest group comprised amateur pianists (42.4%); 16.5% were amateur musicians with other main instruments, and 18.3% were nonmusicians (for group sizes, see Figure 3 and Table 1). This subgrouping was more specific than in Comeau et al. (2017), who divided the analysis into only two groups: 1) musicians (professional musicians, teachers, and students; $n = 26$ with 17 of them playing the piano as their primary instrument); and 2) nonmusicians (adults who did "not currently play an instrument and who [. . .] never had private music lessons"; $n = 25$; Comeau et al., 2017, p. 201). Comeau et al.'s first group corresponded to the four groups of professional pianists, other professional musicians, amateur pianists, and other amateur musicians in the present study, and their second group to nonmusicians. Musicianship significantly covaried with the language version of the survey (see Table 1; χ^2-Test: $\chi^2 = 33.141 > \chi^2(0.95; 4) = 9.49$, $p < .001$): The German speaking sample comprised a larger proportion of professional musicians and of nonpianist amateur musicians.

TABLE 1. *Contingency Table of Musical Expertise in Both Subsamples*

	Language				
	Chinese		German		
Group	*n*	%	*n*	%	Total *n*
Professional pianists	6	5.5	27	16.1	33
Other professional musicians	5	4.5	25	14.9	30
Amateur pianists	58	52.7	60	35.7	118
Other amateur musicians	10	9.1	36	21.4	46
Nonmusicians	31	28.2	20	11.9	51
Total	110	100	168	100	278

Piano Repertoire Quiz

To control for familiarity with the pieces, we gave the participants a piano repertoire quiz (PRQ) with questions about composers and titles of the stimuli after the main task (for the exact wording, see Table D3 in the online Supplementary Material at mp.ucpress.edu). For the PRQ, audio examples from the three different musical pieces used in the main task (see Subsection Materials and Stimuli, Table 9, and Table D1 in the online Supplementary Material for the list of pieces) were repeated (Stimulus 1 for the Chopin Étude, Stimuli 3 and 5 for the Beethoven Piano Sonata, and Stimuli 7 and 10 for the Mozart Piano Concerto). For each of these three pieces, participants were asked whether they knew: 1) the composer, and 2) the title of the piece, which leads to a maximum score of $3 * 2 = 6$ points. For each composer and title, five responses categories (one of them being "I don't know") were given (see Table D3 in the online Supplementary Material). An overall score was calculated as the sum of correct responses to the six questions. On average, participants answered 2.74 ($SD = 2.24$) of the six questions correctly, suggesting that they were moderately familiar with the pieces (see Table 2). It should be considered that randomly selecting any of the four response options (except "I don't know") would yield a total score of 1.5 correct answers on average.

TABLE 2. *Descriptive Statistics of the Number of Correct Responses in the Piano Repertoire Quiz and Scores From the General Factor of the Gold-MSI (General Musical Sophistication)*

	Correct responses in the PRQ	General Musical Sophistication
M	2.74	86.48
SD	2.24	19.78
Minimum	0	33
Maximum	6	121
Range of possible values	[0, 6]	[18, 126]

Musical Sophistication: Gold-MSI

To control for the putative influence of musical skills and expertise, we assessed the degree of musical sophistication using the general factor of the German version (Schaal et al., 2014) and the Chinese version (Lin et al., 2021) of the Gold-MSI (Müllensiefen et al., 2014). Traditional Chinese speaking participants showed an average Gold-MSI of 80.0 (SD = 17.9, n = 110), which corresponds to the 46th percentile of the underlying Taiwanese norm sample (Lin et al., 2021). The mean of the general factor of the Gold-MSI of the German speaking participants was 90.7 (SD = 19.9, n = 168). This corresponds approximately to the 80th to 85th percentile of the German norm sample (Schaal et al., 2014, p. 445). Comparisons with the possible range of the Gold-MSI (18 to 126) and the quartiles of the norm sample showed that the general degree of musical sophistication in our sample was above average. Overall results of the Gold-MSI questionnaire are shown in Table 2.

Influence of Sound Transmission Features

Information about the quality and characteristics of participants' sound transmission devices such as headphones and loudspeakers was obtained by means of the Headphones and Loudspeaker Test (HALT; Wycisk et al., 2022; Wycisk et al., 2018). Headphones were used by about 35%, and loudspeakers by 65% of participants (see Table D5 of the online Supplementary Material for exact numbers). Most participants (74%) listened to the audio examples in stereo with correct polarity (i.e., the right channel on the right side and vice versa). The remaining participants used switched stereo channels, listened monophonic, or answered the question of HALT incorrectly (see Table D4 of the online Supplementary Material for details). Differences concerning the discrimination abilities of participants with different devices are examined and reported in the Results section.

OPERATIONALIZATION AND DESIGN

Participants in both the original and the replication study had to decide whether a stimulus was performed by a child prodigy or an adult professional pianist. In our study, the measurement of this skill was based on Signal Detection Theory (SDT). SDT relies on the classification of each response to one of the four following categories depending on the actual type of stimulus (prodigy or adult) and the response of the participant (prodigy or adult): 1) hit, 2) miss, 3) false alarm, or 4) correct rejection. The following labeling of the four possible response categories was used (see Table 3): If the stimulus was played by a prodigy, and the participant's correct answer to the question was "prodigy," this was labeled as *hit*. If the answer was erroneously "adult," this was called *false rejection* or *miss*. If a stimulus played by an adult was presented and the answer was "prodigy," this answer was wrong and counted as *false alarm*. Inversely, if the correct answer was "adult," this was considered a *correct rejection* (Macmillan & Creelman, 2005).

Based on a participant's overall proportion of hits and false alarms for the total set of stimuli, the sensitivity index *d'* (*d prime*) was calculated as a measure of the discrimination performance (for calculations of *d'*, see Section B of the online Supplementary Material). The higher *d'*, the better a person could discriminate between child prodigies and professional adult pianists. A *d'* of zero indicates discrimination at chance level. A second indicator according to SDT is a person's response bias *c* (for calculations of *c*, see Section B of the online Supplementary Material). This value describes the tendency to prefer one of the two answering options independent from the correct response category (Macmillan & Creelman, 2005, p. 27).

Treating a prodigy stimulus as *A* and a stimulus by an adult professional musician as *Not A* and labeling the

TABLE 3. *Classification of a Response According to the Four SDT Categories*

		Participant response	
		prodigy	adult professional musician
Stimulus	prodigy	hit	miss
	adult professional musician	false alarm	correct rejection

Note. Grey shaded cells indicate wrong answers. The musical source and performer category is indicated in row Stimulus. Participants' response categories are indicated in column Participant response.

answers "prodigy" and "adult" as "A" and "Not A," respectively, the design corresponded to the so-called *A-Not A* method according to the nomenclature classified by Bi (2015, p. 76) and explained by Düvel and Kopiez (2022). In this design, stimuli were not evaluated by direct comparison (as was the case in Comeau et al.'s [2017] comparison task) but independently. The study design is denoted as a *Replicated Paired A-Not A Design* (Bi & Ennis, 2001, p. 216; Düvel & Kopiez, 2022, p. 6). Using the above-mentioned measures and classifications, we could reformulate the empirical hypotheses, in particular the first one, into statistical hypotheses:

A. H_0: *P(hit) = P(false alarm)* vs. H_1: *P(hit)* $\neq$ *P(false alarm)*
B. $d'_{musicians} > d'_{nonmusicians}$
C. $d'_{professional\ pianists} > d'_{others}$
D. H_0: cor(*d'*, *Gold-MSI*) ≤ 0 vs. H_1: cor(*d'*, *Gold-MSI*) > 0 ("cor" indicating correlation).

MATERIALS AND STIMULI

The stimuli used in this study were extracted from the classical repertoire and were composed by Beethoven, Chopin, and Mozart, respectively. All pieces are also on the list of stimuli of the original study. Our stimuli were extracted from 10 professional CD recordings: five recordings from two piano prodigies (Helen Huang and Umi Garrett, both age 12 at the time of recording) and five corresponding recordings from professional adults (Valentina Lisitsa, Ian Parker, and Sebastian Forster; see Table 9 for the list of pieces). Umi Garrett was chosen as an exemplary prodigy because her recordings, like those of Lisitsa and Parker, had also been selected by Comeau et al. (2017). Helen Huang was considered a prodigy because she had been included in the list of musical prodigies in Wikipedia ("List of Child Music Prodigies," 2018) and was designated as a wunderkind, for example, in the booklet of her first album (Mozart & van Beethoven, 1995). All the selected adult musicians have international reputations and have performed with internationally well-known orchestras, such as the London Symphony Orchestra (Lisitsa and Parker) and the Concertgebouw Chamber Orchestra (Forster). In the original study, Comeau et al. (2017) used recordings of Lisitsa and Parker as well. Therefore, we concluded that the adult pianists have a high level of expertise. As Huang's album is a live recording, we selected a live recording as the matching stimulus from the adult group. Selected passages and additional information are listed in Table D1 in the online Supplementary Material.

According to Feldman's definition, the chosen prodigies Helen Huang and Umi Garrett would have been too old to be classified as prodigies. However, as explained in the introductory section of this article, Shavinina (2016) recommends a slightly higher age threshold: As the two prodigies in the present study are girls, the threshold of 10 to 11 years would apply. Both children were 12 years old at the time of the recording. Therefore, they could be expected to be more experienced and trained in playing the piano than 1 or 2 years before the recording. Thus, if differences between the performances of prodigies and adult professional musicians were to be perceived in this study, we would expect the differences to be even more pronounced in the case of younger prodigies. Comeau et al.'s (2017) results also showed that younger prodigies were easier to distinguish from professionals than were older prodigies.

The respective passages were selected by piano experts (lecturers from the Hanover University of Music, Drama and Media). These experts received a recording of the piece and marked those passages with a length of about 30 s each in the score, which, in their opinion, were representative of the prodigy's ability in terms of technique and musical expression. In contrast to the original study, in which the 165 stimuli were 42 s long on average (SD = 15.4, ranging from 15 to 128), the excerpts of this replication study had a more homogeneous length of 32 s on average (SD = 7, ranging from 20 to 40). In addition, the sound quality of our stimuli was at a professional level as only CD recordings and no concert recordings were considered, contrary to the original study, in which YouTube recordings were used among other sources. To avoid confounding by generation and long-term expertization effects (for an explanation of these terms, see the introduction of this paper), we compared only prodigies from the last generation (born between 1982 and 2000) with today's professional musicians while keeping the difference between adults and child prodigies in our study less than 30 years of age at the time of recording. Of course, there is no official limit for the maximum acceptable age difference for this kind of comparison. However, after internal discussion with piano experts, the age difference of 30 years seemed to be a reasonable limit for our decision on inclusion.

PROCEDURE

The online questionnaire was implemented and presented on the platform SoSci Survey (https://www.soscisurvey.de/) in German and Traditional Chinese. Participants were informed about the objectives and the procedure of the survey and gave informed consent online. Anonymity

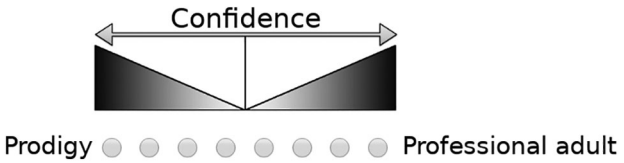

FIGURE 1. Response scale for participant's classification of an audio example (English translation). The question was: "Was the audio sample played by a child prodigy or a professional adult?" The German original version can be found in Figure E1 in the online Supplementary Material at mp.ucpress.edu.

and confidentiality for data handling was ensured. After a short introduction, HALT (Wycisk et al., 2022; Wycisk et al., 2018) was conducted. The short test was split into two parts: The counting of tone events for the adjustment of volume was conducted first, followed by the discrimination task.

For the main task, participants listened to one stimulus at a time (A-Not A design) and decided whether it was played by a child prodigy or an adult professional by ticking one of the options on an 8-point scale (see Figure 1 in this paper and Figure E1 in the online Supplementary Material for the original German version; the same scale was used in Düvel et al., 2020). Simultaneously, the rating scale included participants' rating confidence; that is, how sure they were about their decision.

The rating scheme was explained at the beginning of the test trials. No audio example as a sample was given. Participants could listen to each stimulus repeatedly. All 10 stimuli were randomized with the only restriction being that the same excerpt of a piece was not played twice in immediate succession. At the end of the 10 stimuli, two stimuli (one played by a child prodigy and one by an adult performer) were randomly repeated to control for test–retest reliability of responses. Afterwards, we asked the participants to specify their decision criteria in a free text entry. The explorative analysis of responses by means of a word cloud are displayed in Figure E2 in the online Supplementary Material. Unlike the original study by Comeau et al. (2017), only an identification task and not a comparison task was presented. In a real-life listening situation (e.g., a life concert), no direct comparison would be possible. Therefore, in this context, the identification task (A-Not A design) provided better ecological validity than Comeau et al.'s (2017) comparison task (two-alternative forced-choice design).

At the end of the stimulus identification, participants completed the PRQ and the general factor of the Gold-MSI (Goldsmith Musical Sophistication Index; Lin et al., 2021; Müllensiefen et al., 2014; Schaal et al., 2014). Finally, questions were posed about the participants' musical background and demographics. These included questions about the principal instrument, the musical profession, and the piano expertise. The exact wording of the PRQ and the questions on the musical background can be found in Tables D2 and D3 (see the online Supplementary Material). After this, participants continued with the second part of HALT (Wycisk et al., 2022; Wycisk et al., 2018): the identification of headphone or loudspeaker usage by a spatial location task (location of sounds in the head or around the listener); and monophonic vs. stereophonic listening (counting of tone events on the right ear only). In the end, participants received information about the result of the discrimination task and PRQ. No reimbursement was paid, but the opportunity to take part in a sweepstake was offered (with separate storage of the email address). The entire procedure lasted on average 23.9 min (SD = 4.4).

Results

RESPONSE RATES, SENSITIVITY, BIAS, AND AUC

Overall Results for Total Sample

All statistical analyses for this section were based on JASP (JASP Team, 2020) and R (Version 4.0.3; R Core Team, 2021). The relative frequency was 51.01% for hits and 48.99% for misses (see Table 3 & 4). For correct rejections and false alarms, it was 56.40% and 43.60%, respectively (see Table 3 & 4). Thus, the probability of hits, the probability of correct rejections, and the overall probability of correct responses (53.71%) was higher than chance level (50%). The latter was lower than the proportion of overall correct responses in the original study (57.8%; see last column of Table 4). The mean sensitivity d' was 0.20 (95% CI [0.10, 0.29]; see Table 4 and Figure 2A). This value lay within Bi's (2015, p. 44) benchmarks for a small effect ($0.0 < d' < 0.74$), with 0 corresponding to chance level. The mean discrimination performance in the present study was higher compared to Comeau et al.'s (2017) finding (approximately $d' = 0.13$ for prodigies aged between 11 and 14; see last column of Table 4), but the difference was nonsignificant (one-sample t-test with two-tailed hypothesis: $t(277) = 1.36$, $p = .174$, Cohen's $d = 0.241$ with 95% CI [–0.163, 0.319]). In our study, Bias c was 0.07 (95% CI [0.03, 0.12]; see Table 4 and Figure 2B), which corresponded to a very small bias. In other words, participants showed a weak tendency to classify stimuli as being played by an adult. To consider the confidence ratings, we analyzed the area under the receiver operating characteristic (ROC) curve. An ROC plot shows

TABLE 4. *Descriptive Statistics of the Overall Proportions of Hits, Misses, Correct Rejections, False Alarms, Sensitivity, Response Bias, and Area under Curve for the Total Sample (*N *= 278)*

	M	*SD*	Min.	Max.	*SE*	95% CI		Original study
						LL	*UL*	
Proportion in % for								
Hits (correct responses for prodigy stimuli)	51.01	20.28	0	100	1.22	48.61	53.40	–
Misses (false responses for prodigy stimuli)	48.99	20.28	0	100	1.22	46.60	51.39	–
Correct rejections (correct responses for adult stimuli)	56.40	22.19	0	100	1.33	53.78	59.02	–
False alarms (false responses for adult stimuli)	43.60	22.19	0	100	1.33	40.98	46.22	–
Overall correct responses (for prodigy and adult stimuli)	53.71	15.04	10	90	0.90	51.93	55.48	~54.6
Sensitivity *d'*	0.20	0.82	−2.12	2.12	0.049	0.10	0.29	~0.13
Bias *c*	0.07	0.41	−1.06	1.28	0.024	0.03	0.12	–
Area under curve (AUC)	.54	.18	.00	.96	.011	.51	.56	–

Note. CI = confidence interval; *LL* = lower limit; *UL* = upper limit. The last column shows the results of the identification task of prodigies aged 11 to 14 in the original study by Comeau et al. (2017).

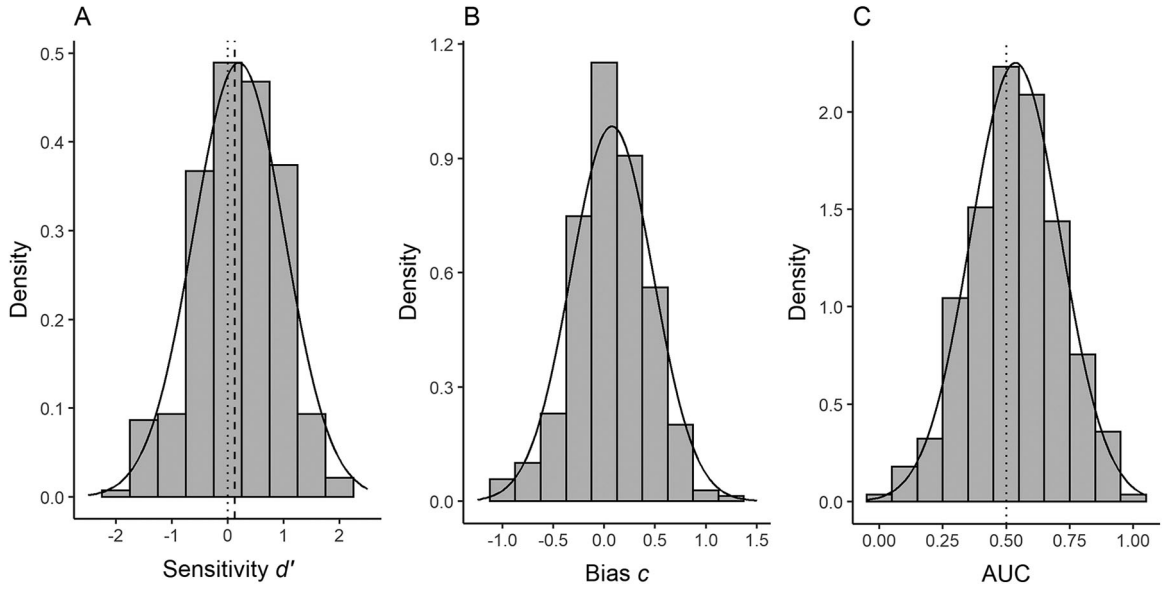

FIGURE 2. Results from the auditory discrimination task between stimuli played by musical prodigies or adult professional pianists: Histogram and density of the normal distribution of (A) Mean sensitivity *d'*, (B) Bias *c*, and (C) AUC (*N* = 278). Dotted vertical lines in (A) and (C) indicate chance levels. The dashed line in (A) indicates sensitivity *d'* in the original study by (Comeau et al. (2017).

"the performance of a binary classification method" and hence, the area under this curve (AUC) "measures the performance of a classifier"—the higher the AUC, the better the classifier (Robin et al., 2011). We calculated the AUC for each participant with a mean AUC of .54 (*SD* = .18; see Table 4 and Figure 2C). An AUC value of .5 is equivalent to classifying by chance. In our study, the mean AUC was significantly higher than chance level (one-sample *t*-test with one-tailed hypothesis: $t(277) = 3.34$, $p < .001$, Cohen's $d = 0.200$ with 95% CI [-0.037, 0.437]). This difference signified a small effect according to Cohen's benchmarks (Ellis, 2010, p. 41).

Hypothesis A (H_0: P(hit) = P(false alarm)*).* In the next step, McNemar's test was calculated to evaluate Hypothesis A and to decide whether participants responded in a different way to stimuli performed by a prodigy than to stimuli played by an adult professional musician. The null hypothesis was that the proportion of hits (*P(hit)* $= p_a + p_c$) would not differ from the proportion of false alarms (*P(false alarm)* $= p_a + p_b$), and participants would

TABLE 5. *Classification and Frequencies of Response Patterns for Each Pair of Stimuli*

Stimulus		Prodigy		
	Response	"Prodigy"	"Adult musician"	Sums of rows
Adult musician	"Prodigy"	$p_{11} = \frac{a}{\widetilde{N}} = \frac{310}{1390} = 22.3\%$	$p_{12} = \frac{b}{\widetilde{N}} = \frac{296}{1390} = 21.3\%$	$\frac{a+b}{\widetilde{N}} = \frac{606}{1390} = 43.6\% = P(\textit{false alarm})$
	"Adult musician"	$p_{21} = \frac{c}{\widetilde{N}} = \frac{399}{1390} = 28.7\%$	$p_{22} = \frac{d}{\widetilde{N}} = \frac{385}{1390} = 27.7\%$	$\frac{c+d}{\widetilde{N}} = \frac{784}{1390} = 56.4\% = P(\textit{correct rejection})$
Sums of columns		$\frac{a+c}{\widetilde{N}} = \frac{709}{1390} = 51.0\% = P(\textit{hit})$	$\frac{b+d}{\widetilde{N}} = \frac{681}{1390} = 49.0\% = P(\textit{miss})$	100%

Note: For the meaning of *a*, *b*, *c*, and *d*, see Table B1 in the online supplementary material. $\widetilde{N} = k * N = 1390$ was the total number of responses to pairs of stimuli (effective sample size) where $k = 5$ was the number of pairs of stimuli and $N = 278$ was the number of participants.

therefore classify stimuli at chance level (H_0: $p_a + p_c = p_a + p_b$; for the meaning of *a*, *b*, and *c,* see Table B1 in the online Supplementary Material). The alternative hypothesis was that the two proportions would differ (H_1: $p_a + p_c \neq p_a + p_b$; note that the frequency of response pattern *c* is not equal to bias *c*). Table 5 shows the classification of response patterns for each pair of stimuli according to Table 4.5 in Bi (2015, p. 76; see also Table B1 in the online Supplementary Material)

The proportion of hits vs. false alarms was significantly different, as tested by McNemar's test ($X^2_M(1) = \frac{(b-c)^2}{b+c} = 15.27$, $p < .001$; see Düvel & Kopiez, 2022, for an explanation and a discussion). Additionally, an exact McNemar's test using the binomial distribution also showed a significant result ($p < .001$), and thus the null hypothesis (Hypothesis A) could be rejected. The effect size (odds ratio *o*) was estimated as $o = b/c = 0.742$; 95% CI [0.636, 0.864] (see Equation 4.4.14 in Bi, 2015, p. 78). Thus, odds ratio *o* was not equal to 1 (which would mean that participants classified the stimuli according to chance level). To conclude, the hit rate was higher than the false alarm rate, and participants showed a task performance above chance level. Converting $\chi^2_M = 15.27$ to the effect size of $w = 0.105$ for McNemar's test (Bühner & Ziegler, 2009, p. 313), we found a small effect according to Cohen's benchmarks (Ellis, 2010, p. 41). In other words, these results showed that participants did notice a difference between the same musical stimuli performed by prodigies and adult professional musicians.

Differences in Sensitivity Depending on Musical Expertise

The descriptive statistics of sensitivity *d'* for the individual groups are shown in Table 6 and in Figure 3. However, no significant differences in sensitivity between the five different subgroups were found, as tested by an ANOVA, $F(4, 273) = 0.40$, $p = .81$, $\eta^2 = .006$. A more detailed analysis of the relationship between musical expertise (especially regarding piano expertise) was carried out in the analysis of Hypotheses B, C, and D. For Hypotheses B and C, we did not only consider sensitivity *d'* but also the proportion of correct responses to allow for comparisons with Comeau et al.'s (2017) result.

Hypothesis B ($d'_{\text{musicians}} > d'_{\text{nonmusicians}}$). The proportion of overall correct responses was 53.95% (*SD* = 15.38) for musicians (results pooled for the four groups of professional pianists, other professional musicians, amateur pianists, and other amateur musicians: $n = 227$, sensitivity $d' = 0.21$), and 52.55% (*SD* = 13.54) for nonmusicians ($n = 51$, sensitivity $d' = 0.14$; see Table 6). Compared to Comeau et al.'s (2017, p. 203) identification task with prodigies aged between 11 and 14 years, this result was lower for musicians (approximately 59.5% in the original study) and higher for nonmusicians (approximately 49.5% in the original study; see the last column of Table 6). A planned contrast analysis (Kirk, 2013, pp. 154, 176; type difference) did not reveal a significantly higher discrimination rate (as measured by sensitivity *d'*) for the pooled results when comparing the musician and nonmusician groups ($\psi_1 = -0.10$, $t = -0.74$, $SE = 0.13$, $df = 273$, $p = .46$). Thus, Hypothesis B was rejected.

Hypothesis C ($d'_{\text{professional pianists}} > d'_{\text{others}}$). For Hypothesis C, we examined whether professional pianists had a better discrimination performance than other participants. A planned contrast analysis (type: Helmert, $\psi_2 = 0.12$, $t = 0.79$, $SE = 0.15$, $df = 273$, $p = .43$) suggested no significant differences between professional pianists on the one hand ($n = 33$; mean proportion of overall correct responses = 55.76%, *SD* = 18.03, sensitivity $d' = 0.31$) and

TABLE 6. *Descriptive Statistics for the Discrimination Performance of Groups of Different Musical (Piano) Expertise*

			Sensitivity *d'*						Correct responses (%)		
						95% CI					
Musical (piano) expertise	*n*	%	*M*	*SD*	*SE*	*LL*	*UL*	Original study *M*	*M*	*SD*	Original study *M*
Professional pianists	33	11.9	0.31	1.00		−0.045	0.665	–	55.8	18.03	–
Other professional musicians	30	10.8	0.30	0.79		0.005	0.595	–	55.7	14.55	–
Amateur pianists	118	42.4	0.18	0.82		0.031	0.329	–	53.5	15.27	–
Other amateur musicians	46	16.5	0.15	0.79		−0.085	0.385	–	52.8	14.40	–
Nonmusicians	51	18.3	0.14	0.74		−0.068	0.348	–	52.5	13.54	–
Total	278	100	0.20	0.82		0.103	0.297	~0.13	53.7	15.04	~54.6
Hypothesis B											
Musicians	227	81.7	0.21	0.83	0.06	0.101	0.318	~0.30	53.96	15.38	59.5
Nonmusicians	51	18.3	0.14	0.74	0.10	−0.068	0.348	~−0.05	52.55	13.54	49.5
Hypothesis C											
Professional pianists	33	11.9	0.31	0.98	0.17	−0.070	0.349	–	55.76	18.03	–
Others	245	88.1	0.18	0.79	0.05	−0.032	0.660	–	53.43	14.62	–

Note. CI = confidence interval; *LL* = lower limit; *UL* = upper limit. The last column shows the results of the identification task of prodigies aged 11 to 14 in the original study by Comeau et al. (2017).

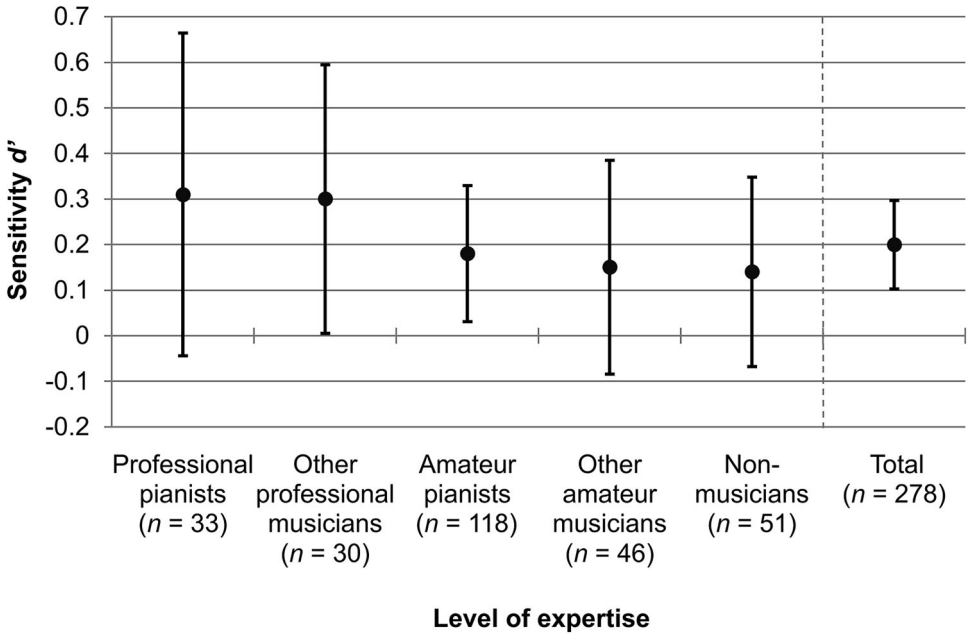

FIGURE 3. Discrimination performance (*d'*) of groups with different musical expertise. Error bars indicate 95% confidence intervals.

other professional musicians, amateur pianists, other amateur musicians, and nonmusicians on the other (n = 245; mean proportion of overall correct responses = 53.43%, SD = 14.62, sensitivity d' = 0.18; see Table 6). Therefore, Hypothesis C was rejected.

*Hypothesis D (cor(*d', *Gold-MSI) > 0).* By testing Hypothesis D, we wanted to investigate a possible correlation between general musical sophistication and discrimination performance. The general factor of the Gold-MSI and sensitivity d' did not correlate at

TABLE 7. *Sensitivity of Groups With Different Task Language (N = 278)*

	t-Test			
Variable	*t*	*df*	*p*	Cohen's *d*
Sensitivity *d'*	−0.17	241	.862	−0.021
AUC	−1.43	238	.155	−0.174
Proportion of overall correct responses	−0.14	240	.886	−0.018

a statistically significant level (r = .05, p = .185, one-tailed). Thus, Hypothesis D was rejected.

Piano Repertoire Quiz. The number of correct answers in the PRQ correlated slightly with sensitivity *d'* (Pearson's r = .14, p = .020, two-tailed) and participants' age (Pearson's r = .20, p < .001, two-tailed) and strongly with the Gold-MSI (Pearson's r = .49, p < .001, two-tailed), according to the benchmarks of Ellis (2010, p. 41), ruling out a strong impact of familiarity with the pieces on discrimination performance

Differences in Sensitivity Between Participants With Different Audio Playback Devices

We collected information on the participants' playback devices to control for the influence of their quality on participants' discrimination performance. Therefore, regarding headphones vs. loudspeakers and mono vs. stereophonic playback, we categorized participants into two groups for each of the two variables (headphones vs. loudspeakers, stereophonic vs. monophonic playback; see Table D5 in the online Supplementary Material for details). The use of headphones vs. loudspeakers and of mono vs. stereophonic playback did not have a statistically significant impact on participants' discrimination performance (also see Table D5 for details).

Differences in Sensitivity Between Different Languages

The questionnaire was answered by two subgroups in German or Traditional Chinese. The task language may relate to musical culture and, thus, to differences in the amount of time spent listening to Western classical music, which could affect responses. To control for putative differences between participants from different language backgrounds (German and Traditional Chinese), we ran two-tailed Welch Two Sample *t*-tests to compare sensitivity *d'*, the AUC, and the proportion of overall correct responses. However, none of the *t*-tests was significant (see Table 7).

CONFIDENCE RATINGS OF RESPONSES

Participants rated the stimuli on an 8-point rating scale, indicating their classification of the stimuli as played by a prodigy or an adult professional musician, and, simultaneously, evaluated the confidence of their rating (see Figure 1 and Figure E1 in the online Supplementary Material). Therefore, ratings of 1 or 8 represented a high confidence in the participants' judgement, whereas ratings of 4 or 5 (i.e., the center of the scale) represented a high uncertainty. To analyze the confidence without the classification as prodigy or adult stimulus, we transformed the results: Ratings from 1 to 4 were kept (indicating high to low confidence), and ratings from 5 to 8 were mirrored to the range of 4 to 1 (8 is converted to 1 [high confidence], 5 into 4 [low confidence], and the values in between accordingly). Descriptive statistics of confidence ratings averaged across all stimuli, the five prodigy stimuli, or the five adult stimuli are shown in Table 8.

TABLE 8. *Descriptive Statistics of the Confidence Ratings (1 = High Confidence, 4 = low Confidence)*

	M	*SD*	*Min.*	*Max.*	*SE*
Total of 10 stimuli	2.50	0.657	1	4	0.039
Five prodigy stimuli	2.49	0.716	1	4	0.043
Five adult stimuli	2.50	0.680	1	4	0.041

To investigate the relationship between participants' discrimination performance and the confidence of their classification, we calculated correlations: Sensitivity *d'* did not correlate with the confidence averaged over all stimuli (Pearson's r = –.03, p = .63). For the prodigy stimuli (Pearson's r = –.02, p = .71) as well as for the stimuli by adult professionals (Pearson's r = .00, p = .95), confidence did not correlate with the proportion of correct answers. Therefore, no systematic effects of the confidence ratings could be observed.

TEST–RETEST RELIABILITY, INTERNAL CONSISTENCY, AND JUDGEMENT CONSISTENCY

To control for test–retest reliability, one prodigy and one adult stimulus, which were randomly selected for each participant, were presented twice. Correlations between the responses (ratings from 1 to 8) to the same stimulus, which were presented both in the test and in the retest, were calculated. Due to an unexpected technical problem, retests for prodigy as well as for adult stimuli were not collected from the full sample. Overall, the correlation between test and retest scores of the prodigy stimuli was small, as indicated by a Spearman's value of $\rho_{prodigy}$ = .152 (p < .05 for two-tailed hypothesis, n = 175). For the adult stimuli, test–retest correlation was also of small effect size (ρ_{adult} = .183, p < .05 for two-tailed hypothesis, n = 179). As test–retest reliability depends on the length of the test, a retest of just two items is not a representative measure. To estimate the correlation for the full set of items ($\rho_{complete}$), the

TABLE 9. *Relative Frequencies of Correct Responses and AUC for Each of the 10 Stimuli*

No.	Piece	Performer	Correct responses (%)	AUC for stimulus pairs
1	Chopin: Étude, Op. 10 No. 5 in G flat major, "Black	Child	45.3%	.53
2	Keys"	Adult	60.8%	
3	Beethoven: Piano Sonata No. 14 in C sharp minor,	Child	62.2%	.66
4	Op. 27 No. 2, I. Adagio sostenuto	Adult	67.3%	
5	Beethoven: Piano Sonata No. 14 in C sharp minor,	Child	60.8%	.57
6	Op. 27 No. 2, III. Presto agitato	Adult	52.2%	
7	Mozart: Piano Concerto No. 23 in A major, K. 488,	Child	38.8%	.45
8	Allegro (Live)[1]	Adult	54.0%	
9	Mozart: Piano Concerto No. 23 in A major, K. 488,	Child	47.8%	.49
10	Allegro (Live)[1]	Adult	47.8%	

Note. [1]Two different excerpts were chosen from the same piece. For more detailed information on the pieces and excerpts, see Table D1 in the online Supplementary Material.

Spearman–Brown prophecy formula was applied (Revelle & Condon, 2018, p. 721), resulting in a corrected correlation value of $\rho_{complete,prodigy} = .47$ for the prodigy stimuli and $\rho_{complete,adult} = .53$ for the stimuli of adult professional musicians.

As a measure of internal consistency, Cronbach's α was calculated for stimuli played by a prodigy and for stimuli played by an adult professional musician. Analyses resulted in $\alpha_{prodigy} = -.135$ and $\alpha_{adult} = .096$. A negative value for $\alpha_{prodigy}$ resulted from the negative correlation of the rating of stimulus 1 with the total scale. To conclude, both scales showed no internal consistency.

Another form of judgement consistency was measured analogous to Comeau et al. (2017, p. 205): The one or two test–retest stimuli pairs were considered, and the percentage of similarly judged pairs was calculated. Answers were regarded as similar if a stimulus pair was judged "prodigy" twice or "adult" twice. For example, if a participant mistakenly assigned the same prodigy recording first to an adult and then correctly to a prodigy in the retest and subsequently correctly assigned another recording by an adult twice to a professional pianist in the test and the retest, the consistency was 50%. As mentioned above, not all participants completed two retests: 10.1% (n = 28) completed none of the retests, 52.5% (n = 146) completed one, and 37.4% (n = 104) of the participants completed two retest items. Judgement consistency was calculated for participants taking one or two retest items (n = 250). The mean consistency between test and retest was 56.2% (SD = 44.5). This exceeded the consistency value resulting from chance level (50%). The difference was within the range of a small effect size (one-sample t-test with one-tailed hypothesis: $t(249) = 2.20$, $p = .014$, Cohen's $d = 0.139$ with 90% CI [0.015, 0.264]).

Overall, test–retest reliability was low (around .5); the measurements showed nearly no internal consistency, and the judgment consistency was slightly above chance level. These measures can be interpreted as indicators of the demanding task difficulty.

ANALYSIS ON THE ITEM-LEVEL

The proportion of correct responses to the 10 different stimuli varied between 38.8% (Item 7) and 67.3% (Item 4). The relative frequencies of correct responses for all 10 items are shown in Table 9. As we did for each participant, we calculated the AUC for each pair of stimuli (see Table 9). Stimuli 3 and 4 showed the highest values of AUC indicating that they were the easiest items to distinguish between adult and prodigy. This result was confirmed by the analysis of the latent variable discriminability behavior based on Item Response Theory (see Section C and Table D7 in the online Supplementary Material).

In addition to the analysis of item difficulties based on classical test theory, we also conducted a probabilistic analysis of the latent variable discriminability behavior based on Item Response Theory (for results see Section C and Table D7 in the Supplemental Material).

In line with Comeau, we also measured various acoustic features of the stimuli to explore a possible acoustical basis for the judgments (see Table D6 and Figure E3 in the online Supplementary Material). We conducted a regression analysis to predict participant accuracy as measured by the overall percentage of correct responses using the following acoustic features of the stimuli: Mean and *SD* of intensity (in dB; using Praat V 6.0.46) as well as Mean and *SD* of tempo (using MIR Toolbox V 1.7.2 with frame size = 4 s, hop size = 10%). Results of the regression analysis were nonsignificant

(R^2 = .59, F[4, 5] = .76, p = .273; see Table D8 in the online Supplementary Material). Therefore, no model predicting the percentage of correct responses by using the acoustic measures could be determined. Nevertheless, some correlations could be identified (see Figure E3 in the online Supplementary Material)—but due to the very small sample size (n = 10) they should only be seen as a first explorative approach to possible relations. To elucidate what cues participants used to make their judgments, we investigated whether acoustic features differed between stimuli that tended to be identified as prodigy or as professional adult. As shown in Figure E4 (C) in the online Supplementary Material, most stimuli that were classified as being played by an adult were slower than those which were classified as being played by a child prodigy. Such a tendency is not discernible for the remaining acoustic features.

As the data of the 10 stimuli was paired—five different excerpts of pieces played by two different interpreters (prodigy and professional adult)—we also analyzed the pairs of stimuli by calculating the absolute difference of each of the four acoustic features for each pair of stimuli. Although we found medium to large, but nonsignificant, correlations with the average proportion of correct responses as well as with the AUC of the stimulus pairs, it is clear that changing a single data point could have resulted in a very different correlation coefficient (e.g., large negative to large positive). Based on this very small sample size of n = 5, no general conclusion—not even assumptions of possible relations—can be made.

Discussion

"Deciding whether or not a child is a prodigy is not as easy as it might appear to be" (Howe, 2000, p. 312). Starting from this statement, in this replication study we investigated whether: a) listeners could discriminate between musical performances of child prodigies and adult professionals, b) professional and amateur musicians could discriminate between prodigies and adult professionals better than nonmusicians, c) professional pianists showed the highest discrimination performance, and d) there was a correlation between discrimination ability and the degree of musical sophistication as measured by the Gold-MSI. In contrast to the original study by Comeau et al. (2017), we were able to reduce methodological deficiencies such as confounding influences by different recording qualities, different generations of pianists, a very long participation time, and a strongly varying length of stimuli. As was the case in Comeau et al.'s (2017) study, participants were able to discriminate between child prodigies and adult professionals by listening to comparable music recordings. These results are even observed in children aged 12, so we assume that the difference is in fact more pronounced in younger child prodigies. The rate at which listeners could distinguish prodigies from adults was slightly above chance. This means that one of Comeau et al.'s (2017) results could be successfully replicated. However, this main result conflicts with Feldman's (1993, p. 188) definition of a prodigy as a child under 10 years who gives a musical performance on the level of an adult professional musician and whose performance is not likely to be distinguished from that of a professional musician. Therefore, the outcome of this study questions this definition of a musical prodigy on the fundamental level of empirical data. Furthermore, the maximum age of 10 years was probably set arbitrarily by Feldman and should be questioned in future research and in terms of defining a child prodigy. The results of this study lead to the conclusion that either the young pianists used in this study did not belong to the category of musical prodigies or that Feldman's definition might be inconsistent and unfeasible for operationalization. Other criteria, such as playing technique and musical expression, pace of development, and so on, might be necessary for a conclusive definition of a child prodigy.

In Comeau et al.'s (2017) original study, "musicians generally performed better than nonmusicians" (p. 203). Musicians and pianists in our replication study did not perform significantly better than the other groups. Furthermore, we did not find any relation between musical skills and expertise as measured by the Gold-MSI. This suggests that musical expertise does not play an important role when it comes to recognizing child prodigies by their musical recordings. Thus, this result of Comeau et al. (2017)—namely, of musicians outperforming nonmusicians—could not be replicated.

We wanted to know what indicators participants would use to decide by whom the audio was played. For this reason, we asked them to specify the criteria they applied (see Figure E2 in the online Supplementary Material). One of the "criteria" which was often named was "gut instinct" ("Bauchgefühl" in German). From this, it can be concluded that in sum the discrimination task remained difficult.

LIMITATIONS AND FUTURE PERSPECTIVES

We are aware that our sample is not representative of the general population, because 68.3% of participants played the piano. Nevertheless, musical expertise did not seem to have an effect on participants' discrimination performance. Therefore, we assume that the findings can be transferred to a general population.

A main difficulty of this study was to find appropriate recordings of musical prodigies. Unlike Comeau et al. (2017), we aimed for high audio quality of all recordings. As very few prodigies have possibilities for professional recordings and CD productions, it was hard to find a collection of adequate stimuli which had also been used in the original study. For this reason, we used a smaller set of stimuli with fewer individual musicians and fewer trials overall than the original study. On the one hand, this could lead to lower reliability and generalizability, but, on the other hand, it kept the duration of the testing procedure at a tolerable level and, therefore, ruled out a possible confounding effect caused by fatigue. Furthermore, studio recordings are often edited to produce the best results possible for the artist. However, this might not fully reflect the prodigy's actual level of musical ability, and the use of studio recordings might level out existing differences between performer groups. Considering this, we suggest the future use and production of high-quality live performances as the ideal stimuli for this kind of studies. For this condition, we predict a much higher discriminability value. For future research considering a broader range of musicians, pieces, and a larger number of stimuli from live recordings, researchers might consider the use of professional recordings by engaging current musical prodigies as well as professional adult musicians. In this case, the selection of pieces and passages would also be much easier to control. In addition to a quantitative analysis, a qualitative design would help us gain more information on the cues people use to perform the classification task. A final question would be at which age the differences between performer groups will converge. An answer can only be given by a long-term study and repeated live recordings of a standardized repertoire. We assume that this point will be above the age of 10. It could be the basis for an empirical definition of musical child prodigies. As long as these tasks remain uncompleted, we suggest a critical rethinking of Feldman's definition of a prodigy for the musical domain.

Author Note

Supplementary Material is available at mp.ucpress.edu. Data and R scripts are available from https://osf.io/jgwdt/?view_only=8230efa3804a4314bffb1f44907b3bf0. Sound examples can be obtained from the corresponding author upon reasonable request.

This research received no specific grant from any finding agency in the public, commercial, or not-for-profit sectors. A preliminary version of this study was presented as a poster at the 34th annual conference of the German Society for Music Psychology in September 2018 in Gießen, Germany. The authors would like to thank Prof. Wolfgang Zill and Dinara Devisheva for selecting the excerpts for the stimuli, Hsin-Rui Lin for translating the questionnaire into Traditional Chinese, and Yves Wycisk for the implementation of HALT.

Author contributions: VP conceived the study and collected the data; RK conceived the study; VP, ND, and RK analyzed the data and wrote the manuscript.

Correspondence concerning this article should be addressed to Reinhard Kopiez, Hanover University of Music, Drama and Media, Neues Haus 1, 30175 Hanover, Germany. E-mail: reinhard.kopiez@hmtm-hannover.de

References

Barrington, D. (1770). Account of a very remarkable young musician. *Philosophical Transactions*, *60*, 54–64.

Baumgarten, F. (1930). *Wunderkinder: Psychologische Untersuchungen* [Prodigies: Psychological investigations]. Barth.

Bennett, B. M., & Underwood, R. E. (1970). On McNemar's test for the 2 x 2 table and its power function. *Biometrics*, *26*, 339–343.

Bi, J. (2015). *Sensory discrimination tests and measurements: Sensometrics in sensory evaluation* (2nd ed.). Wiley.

Bi, J., & Ennis, D. M. (2001). Statistical models for the A-Not A method. *Journal of Sensory Sturlies*, *16*, 215–237. https://doi.org/10.1111/j.1745-459X.2001.tb00297.x

Bühner, M., & Ziegler, M. (Eds.). (2009). *Statistik für Psychologen und Sozialwissenschaftler* [Statistics for psychologists and social scientists]. Pearson.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Comeau, G., Vuvan, D. T., Picard-Deland, C., & Peretz, I. (2017). Can you tell a prodigy from a professional musician? *Music Perception*, *35*(2), 200–210. https://doi.org/10.1525/MP.2017.35.2.200

Düvel, N., & Kopiez, R. (2022). The paired A-Not A design within signal detection theory: Description, differentiation, power analysis and application. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-021-01728-w

Düvel, N., Kopiez, R., Wolf, A., & Weihe, P. (2020). Confusingly similar: Discerning between hardware guitar amplifier sounds and simulations with the Kemper Profiling Amp. *Music & Science*, *3*, 1–16. https://doi.org/10.1177/2059204320901952

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press.

Feldman, D. H. (1986). *Nature's gambit: Child prodigies and the development of human potential*. Basic Books.

Feldman, D. H. (1993). Child prodigies: A distinctive form of giftedness. *Gifted Child Quarterly*, *37*, 188–193. https://doi.org/10.1177/001698629303700408

Feldman, D. H., & Morelock, M. J. (2011). Prodigies and savants. In R. J. Sternberg & S. B. Kaufman (Eds.), *The Cambridge handbook of intelligence* (pp. 210–234). Cambridge University Press.

Graus, A. (2021). Child prodigies in Paris in the belle époque: Between child stars and psychological subjects. *History of Psychology*, *24*, 255–274. https://doi.org/10.1037/hop0000192

Howe, M. J. A. (2000). Prodigies. In A. E. Kazdin (Ed.), *Encyclopedia of psychology* (Vol. *6*, pp. 312–313). American Psychological Association.

JASP Team (2020). JASP (Version 0.14.1) [Computer software]. https://jasp-stats.org/

Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral sciences* (4th ed.). Sage.

Kopiez, R. (2011). The musical child prodigy (wunderkind) in music history: A historiometric analysis. In I. Deliège & J. W. Davidson (Eds.), *Music and the mind: Essays in honour of John Sloboda* (pp. 225–236). Oxford University Press.

Kopiez, R., & Lehmann, A. C. (2016). Musicological reports on early 20th-century musical prodigies: The beginnings of an objective assessment. In G. McPherson (Ed.), *Musical prodigies: Interpretations from psychology, education, musicology and ethnomusicology* (pp. 169–184). Oxford University Press.

Lehmann, A. C. (2006). Historical increases in expert music performance skills: Optimizing instruments, playing techniques, and training. In E. Altenüller, M. Wiesendanger, & J. Kesselring (Eds.), *Music, motor control and the brain* (pp. 3–22). Oxford University Press.

Leiner, D. (2019). Too fast, too straight, too weird: Post hoc identification of meaningless data in internet surveys. *SSRN Electronic Journal*, *13*, 229–248. https://doi.org/10.18148/srm/2019.v13i3.7403

Lin, H.-R., Kopiez, R., Müllensiefen, D., & Wolf, A. (2021). The Chinese version of the Gold-MSI: Adaptation and validation of an inventory for the measurement of musical sophistication in a Taiwanese sample. *Musicae Scientiae*, *25*, 226–251. https://doi.org/10.1177/1029864919871987

List of child music prodigies. (2018, September 28). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=List_of_child_music_prodigies&oldid=849548952

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Lawrence Erlbaum.

Marion-St-Onge, C., Weiss, M.W., Sharda, M., & Peretz, I. (2021). What makes musical prodigies? *Frontiers in Psychology*, *11*, 1–13, Article 566373. https://doi.org/10.3389/fpsyg.2020.566373

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*(2), 153–157. https://doi.org/10.1007/BF02295996

Miettinen, O. S. (1968). The matched pairs design in the case of all-or-none responses. *Biometrics*, *24*, 339–352.

Mozart, W., & van Beethoven, L. (1995). *Introducing Helen Huang: Mozart: Piano Concerto No. 23, Beethoven: Piano Concerto No. 1* [Album recorded by H. Huang, New York Philharmonics, K. Mazur]. Teldec. (Original works published 1786 and 1798)

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), 1–23, e89642. https://doi.org/10.1371/journal.pone.0089642

Olbertz, F. (2010). Wunderkind. In H. de la Motte-Haber, H. von Loesch, G. Rötter, & C. Utz (Eds.), *Handbuch der systematischen Musikwissenschaft* [Handbook of systematic musicology]: *Vol. 6. Lexikon der systematischen Musikwissenschaft* [Lexicon of systematic musicology] (pp. 528–529). Laaber.

R Core Team (2021). *R: A language and environment for statistical computing* (Version 4.0.3). R Foundation for Statistical Computing [Computer software]. https://www.R-project.org/

Revelle, W., & Condon, D. M. (2018). Reliability. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (Vol. *2*, pp. 709–749). Wiley.

Richet, C. (1900). Un cas remarquable de précocité musicale. *L'année psychologique*, *7*, 657–659. https://www.persee.fr/doc/psy_0003-5033_1900_num_7_1_3281

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(77), 1–8. https://doi.org/10.1186/1471-2105-12-77

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. https://doi.org/10.1177/1029864914541851

Shavinina, L. V. (2016). On the cognitive–developmental theory of the child prodigy phenomenon. In G. E. McPherson (Ed.), *Musical prodigies: Interpratations from psychology, education, musicology, and ethnomusicology* (pp. 259–278). Oxford University Press.

Solomon, A. (2012). Far from the tree: Parents, children and the search for identity. Scribner.

Wycisk, Y., Kopiez, R., Bergner, J., Sander, K., Preihs, S., Peissig, S., & Platz, F. (2022). The headphone and loudspeaker test – part I: Suggestions for controlling characteristics of playback devices in internet experiments. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-022-01859-8

Wycisk, Y., Kopiez, R., & Wolf, A. (2018, 23–28 July). *Control of headphone and loudspeaker characteristics in online experiments* [Poster presentation]. 15th International Conference on Music Perception and Cognition, Graz. https://static.uni-graz.at/fileadmin/veranstaltungen/music-psychology-conference2018/documents/ICMPC15ESCOM10abstractbook.pdf