

Forschungsberichte

Musikalischer g-Faktor oder multiple Faktoren? Struktur und Leistungskennwerte der musikalischen Hörfähigkeit von Jugendlichen

Musical g Factor or Multiple Factors? Structure and Norms of Musical Ability of Adolescents

Viola Pausch¹, Daniel Müllensiefen^{*1,2}, Reinhard Kopiez¹

[1] Musikwissenschaftliches Institut, Hochschule für Musik, Theater und Medien Hannover, Hannover, Deutschland. [2] Goldsmiths College, University of London, London, Großbritannien.

Zusammenfassung

Die Struktur musikalischer Fähigkeiten ist bislang nicht eindeutig identifiziert. Die vorliegende Studie untersucht, ob es analog zu Gardners Konzept der multiplen Intelligenzen mehrere unabhängige Musikalitäten gibt oder analog zu Spearman's Generalfaktor der Intelligenz hingegen nur einen allgemeinen Musikalitätsfaktor. Hierbei wird Musikalität im Rahmen der LongGold-Studie mittels Tests zur Beat- und Intonationswahrnehmung sowie zur Melodie- und Emotionsunterscheidung an einer Stichprobe aus Jugendlichen (N = 1786) zwischen acht und 18 Jahren (M = 11,78, SD = 1,68) in Deutschland und Großbritannien gemessen. Insgesamt zeigen die Testwerte mittlere bis hohe Korrelationen. Der Vergleich zweier latenter Variablenmodelle und eine konfirmatorische Faktorenanalysen geben Hinweise darauf, dass das Modell des allgemeinen g-Faktors der Intelligenz wahrscheinlich ebenfalls für musikalische Fähigkeit konzeptualisiert als einfaktorielles Modell gilt. Für die einzelnen Musikalitätstests werden Normwerte bereitgestellt, die nach Alter, Geschlecht und Ausmaß der musikalischen Ausbildung differenziert sind.

Schlüsselwörter: musikalische Fähigkeit, Musikalität, Gold-MSI-Fragebogen, Jugendliche, empirische Normen

Abstract

The structure of musical abilities has not yet been clearly identified. The present study investigates whether musicality is best modelled as a collection of several independent musicalities, analogous to Gardner's concept of multiple intelligences, or as one general musicality factor, analogous to Spearman's general "g" factor of intelligence. As part of the LongGold study, we measured musicality using tests of beat and intonation perception as well as melodic and emotional discrimination in a sample of adolescents (N = 1786, age range: 8–18 years, M = 11,78, SD = 1,68) from Germany and the United Kingdom. Overall, test scores correlated with medium to high effect sizes. The comparison of several latent variable models and a confirmatory factor analysis suggested that probably, a model similar to the general g-factor model of intelligence also applies to musical listening ability conceptualized as a single-factor model. Norm values are provided for the individual musicality tests differentiated according to age, gender, and the extent of musical training.

Keywords: musical ability, musicality, Gold-MSI questionnaire, adolescents, empirical norms

Jahrbuch Musikpsychologie, 2022, Vol. 30: Musikpsychologie – Empirische Forschungen - Ästhetische Experimente, Artikel e89,
<https://doi.org/10.5964/jbdgm.89>

Eingereicht: 2021-01-21. Akzeptiert: 2022-01-17. Publiziert (VoR): 2022-03-16.

Begutachtet von: Thomas Schäfer; Gudrun Schwarzer.

*Korrespondenzanschrift: Musikwissenschaftliches Institut, Hochschule für Musik, Theater und Medien Hannover, Neues Haus 1, 30175 Hannover, Deutschland. E-Mail: D.Muellensiefen@gold.ac.uk



Dieser Open-Access-Artikel steht unter den Bedingungen einer Creative Commons Namensnennung 4.0 International Lizenz, CC BY 4.0 (<https://creativecommons.org/licenses/by/4.0/deed.de>). Diese erlaubt für beliebige Zwecke (auch kommerzielle) den Artikel zu verbreiten, in jedem Medium zu vervielfältigen, Abwandlungen und Bearbeitungen anzufertigen, unter der Voraussetzung, dass der Originalartikel angemessen zitiert wird.

Ist es möglich, dass eine Person ein gutes Rhythmusgefühl hat, aber beim Singen keinen Ton trifft? Gibt es Personen, die zwar mit Leichtigkeit jede Melodie im Gedächtnis behalten, aber kein Gespür dafür haben,

welche Emotionen ein Musikstück vermittelt? Obwohl bereits seit mehr als 200 Jahren über Musikalität und musikalische Fähigkeiten geforscht wird (siehe z. B. [Michaelis, 1805](#)), gibt es keine definitive Antwort auf diese Fragen, da nicht bekannt ist, wie viele verschiedene, unabhängige Facetten von Musikalität existieren. Die Struktur musikalischer Fähigkeiten ist bislang nicht eindeutig identifiziert ([Müllensiefen & Hemming, 2018](#), S. 115; [Gembris, 1997](#), S. 869) und aus den Forschungsergebnissen bisheriger Studien können „keine endgültigen Aussagen“ ([Gembris et al., 1997](#), S. 869) darüber abgeleitet werden. Angelehnt an [Fodors \(1983\)](#) Theorie der Modularität des Geistes haben [Peretz und Coltheart \(2003\)](#) ein Modell für die Verarbeitung von Musik formuliert, welches in verschiedene Module gegliedert ist. Dieses Modell legt nahe, dass die Verarbeitung von Musik modular in neuronal isolierbaren Komponenten organisiert ist. Als Beleg dafür werden musikbezogene Defizite bei neurologisch beeinträchtigten Menschen aufgeführt. In der hier vorgestellten Studie wird die Frage beantwortet, ob dieses Konzept der Einteilung musikalischer Verarbeitungsprozesse in mehrere Module auch auf musikalische Fähigkeiten im Jugendalter übertragbar ist. Musikalische Fähigkeiten werden hierbei verstanden als geschätztes und nicht direkt beobachtbares Entwicklungspotential, welches musikalischen Fertigkeiten zugrunde liegt. Musikalische Fertigkeiten sind wiederum definiert als tatsächlich erbrachte und über geeignete Tests messbare, lern- und erfahrungsabhängige, musikalische Leistungen ([Müllensiefen & Hemming, 2018](#), S. 97; [Gembris et al., 1997](#), S. 868).

In der Geschichte der Musikalitätstests spiegelt sich die Grundfrage über die Struktur musikalischer Fähigkeiten in den Testkonzeptionen und -konstruktionen von [Wing \(1961\)](#) und [Seashore \(1919b\)](#) wider: Als „der erste moderne standardisierte Musiktest“ ([Müllensiefen & Hemming, 2018](#), S. 95) ermittelt der Test *Seashore Measures of Musical Talent* ([Seashore, 1919b](#)) mittels sechs Subtests für Tonhöhe, Lautstärke, Rhythmus, Tondauer, Timbre und tonales Gedächtnis ([Müllensiefen & Hemming, 2018](#), S. 103) „ein Profil einzelner und unverbundener grundlegender musikalischer Fertigkeiten“ ([Müllensiefen & Hemming, 2018](#), S. 99) gemäß eines multifaktoriellen Musikalitätskonzepts. Die Ergebnisse der einzelnen Subtests werden jedoch nicht zu einem Summenwert zusammengefasst, sondern bilden stattdessen ein Profil multipler, voneinander relativ unabhängiger musikalischer Fähigkeiten ([Boyle & Radocy, 1987](#), S. 143). Dies macht deutlich, dass Seashore ein einheitliches Konstrukt musikalischer Fähigkeiten als eine einzelne Begabung ablehnte. Im Gegensatz dazu stehen [Wings \(1961\) Standardized Tests of Musical Intelligence](#) und sein Konzept des musikalischen General- bzw. *g*-Faktors. Gemäß eines Generalfaktormodells ([Gembris et al., 1997](#), Sp. 869) nimmt er an, dass einzelnen elementaren, „messbaren musikalischen Fertigkeiten ein einheitliches Konstrukt zugrunde liegt“ ([Müllensiefen & Hemming, 2018](#), S. 99) – und kein multiples Konstrukt wie bei Seashore – und dass sie als musikalischer Generalfaktor miteinander verbunden sind. Daher aggregiert der Wing-Test die Ergebnisse der einzelnen sieben Subtests zu einem einzigen Gesamtwert für musikalische Intelligenz, dem sogenannten Musikalitätsquotienten (MQ; [Boyle & Radocy, 1987](#), S. 146), der eine Analogie zum *g*-Faktor in der Intelligenzforschung darstellt.

Eine frühere Studie, die allgemeiner die Faktorenstruktur von Hörfähigkeiten untersuchte, stammt von [Stankov und Horn \(1980\)](#); vgl. [Müllensiefen, 2017b](#), S. 130). In einer Re-Analyse der Testbatterien, unter anderem von [Seashore et al. \(1960\)](#), [Wing \(1962\)](#) und [Drake \(1954\)](#), fanden sie zwischen den Ergebnissen von 241 erwachsenen Männern in vielen der insgesamt 44 untersuchten Subtests starke Korrelationen. Dadurch konnten sie „sieben Faktoren auditorischer Fähigkeiten“ ([Müllensiefen, 2017b](#), S. 130) identifizieren, von denen drei hauptsächlich musikalische Fähigkeiten beschreiben: „auditorische Kognition von Zusammenhängen, Unterscheidung von Soundmustern, Beibehalten und Beurteilen von Rhythmen“ (vgl. [Müllensiefen, 2017b](#), S. 130). Da die Studie von [Stankov und Horn \(1980\)](#) nicht ausschließlich musikalische Wahrnehmungsfähigkeiten betrachtet,

sind die Ergebnisse nicht direkt auf die oben genannte Problemstellung übertragbar. Die vorliegende Arbeit beschränkt sich daher auf musikalische Fähigkeiten und untersucht zudem eine größere Stichprobe ($N = 1786$) als die von Stankov und Horn. Eine Studie von Kidd et al. (2007) verwendete ebenfalls den faktoranalytischen Ansatz und Modellvergleiche, betrachtet jedoch allgemeinere auditorische Fähigkeiten und arbeitete zudem nicht mit musikbezogenen Tests.

Im Rahmen der Langzeitstudie *LongGold* (<https://longgold.org>) zur Entwicklung musikalischer Fähigkeiten im Jugendalter werden musikalische Fähigkeiten in verschiedenen, unabhängig voneinander konstruierten Musikalitätstests zur Beat- und Intonationswahrnehmung sowie zur Melodie- und Emotionsunterscheidung untersucht. Auf dieser Datengrundlage soll in der hier vorgestellten Studie die Frage beantwortet werden, ob das Konzept der Einteilung musikalischer Verarbeitungsprozesse in mehrere Module wie bei Peretz und Coltheart (2003) auch auf musikalische Fähigkeiten im Jugendalter übertragbar ist. Die Herangehensweise ist hierbei „durch die Intelligenzforschung inspiriert“ (Gembris et al., 1997, Sp. 869), da diese bereits ähnliche Überlegungen über die Struktur allgemeiner kognitiver Fähigkeiten und entsprechende Strukturmodelle lieferte, obgleich nach wie vor Uneinigkeit darüber herrscht, ob Intelligenz als einzelne Fähigkeit oder als Repertoire mehrerer heterogener, unabhängiger und unkorrelierter Fähigkeiten zu konzeptualisieren sei (Myers, 2014, S. 400; Neyer & Asendorpf, 2018, S. 154).

Eines der Strukturmodelle ist Gardners (2006) Modell der multiplen Intelligenzen, in dem mehrere Formen von Intelligenz unterschieden werden. Spearman (1904b), Jensen (1989) und Mackintosh (2011) hingegen gehen von einem Generalfaktor der Intelligenz aus, mit dem alle anderen Subfaktoren verbunden sind. Spearmans einfaches Faktormodell wurde über die letzten Jahrzehnte zu hierarchischen Modellen mit mehreren Ebenen ausdifferenziert: zum Beispiel zusammengefasst in der Cattell-Horn-Carroll-Theorie der kognitiven Fähigkeiten (CHC-Theorie; McGrew, 2005). Dabei wird weiterhin der *g*-Faktor als grundlegende mentale Fähigkeit angenommen. Musikalische Fähigkeiten werden in der CHC-Theorie unter der breiten Domäne *Auditory processing* eingeordnet. Dort werden *Musical Discrimination and Judgment* und *Maintaining and Judging Rhythm* als spezifische Fähigkeiten unterschieden. Die Frage, die in unserer Studie beantwortet werden soll, ist, ob diese spezifisch musikalischen Fähigkeiten auch wieder eine hierarchische Struktur aufweisen oder eher unverbunden sind.

Ziel der Studie

Durch eine Faktorenanalyse sollen im ersten Teil dieser Arbeit folgende Leitfragen beantwortet werden: Existiert nur ein einziger allgemeiner Musikalitätsfaktor analog zu Spearmans (1904b) Generalfaktor der Intelligenz, der den Zusammenhang zwischen verschiedenen Musikalitätstests erklären könnte, oder sind sie hingegen – vergleichbar zu Gardners (2006) Konzept der multiplen Intelligenzen – als multiple Musikalitätsfaktoren unverbunden und liegen damit mehrere unabhängige Formen musikalischer Fähigkeiten vor? Ist eines der konkurrierenden Modelle von Gardner und Spearman auf die Struktur musikalischer Fähigkeiten übertragbar?

Da für viele wichtige Musiktests, wie zum Beispiel denen von Seashore (1919a, S. 11) und Gordon (2004, S. 24), Normen existieren, jedoch noch nicht für die in der *LongGold*-Studie eingesetzten Musiktests, ist ein weiteres Ziel, abhängig von Alter, Geschlecht und Ausmaß musikalischer Ausbildung für jeden Test Normwerte aus den empirischen Daten zu erstellen.

Das LongGold-Projekt

Die empirischen Daten, die die Grundlage dieser Arbeit bilden, wurden im Rahmen des Forschungsprojekts *LongGold* erhoben (<https://longgold.org>). *LongGold* ist eine Langzeitstudie, die musikalische Hörfähigkeiten und nicht-verbale Intelligenz in adaptiven Tests, Persönlichkeit durch das Big-Five-Persönlichkeitsinventar sowie Einstellungen, Selbstkonzept, mentales Wohlbefinden, soziale Fähigkeiten, musikalische und nicht-musikalische Freizeitaktivitäten und demographische Variablen erfasst (Müllensiefen & Elvers, 2018). Ziel des Projekts ist es zu verstehen, wie sich musikalische Fähigkeiten im Alter zwischen zehn und 20 Jahren in Zusammenhang mit Intelligenz, sozialen Kompetenzen und Persönlichkeit entwickeln, und ob Transfereffekte zwischen diesen Bereichen oder Folgeeffekte musikalischer Betätigung auf Schulleistung existieren (Müllensiefen & Elvers, 2018). Darüber hinaus sollen Faktoren identifiziert werden, welche die musikalische Entwicklung beeinflussen, fördern oder verursachen können (Müllensiefen & Elvers, 2018; Müllensiefen, 2017a). Die Durchführung der Studie an Schulen wurde von den Ethikkommissionen der Leibniz Universität Hannover und des Goldsmiths College der University of London sowie den Kultusministerien von Baden-Württemberg und Hessen genehmigt.

Methode

Vorgehensweise

In dieser Studie wird das am besten passende Strukturmodell für musikalische Fähigkeiten anhand einer Stichprobe aus 622 britischen Jugendlichen identifiziert. Anschließend wird geprüft, ob dieses Modell seine Gültigkeit für eine größere Stichprobe aus 1164 Kindern der 5. und 6. Klassen mehrerer Schulen in Deutschland behält.

Stichprobe

Im Rahmen der Langzeitstudie *LongGold* wurden im Schuljahr 2018/19 in Großbritannien 622 und in Deutschland 1168 Jugendliche bezüglich ihrer musikalischen Fähigkeiten untersucht. Vier Schüler*innen aus Deutschland wurden wegen einer auditiven Wahrnehmungsstörung ($n = 1$) und dem Tragen von Hörgeräten ($n = 3$) aus der endgültigen Gesamtstichprobe ausgeschlossen, sodass die verbleibende Gruppe $N = 1786$ Kinder und Jugendliche umfasste. Sie waren zwischen acht und 18 Jahre alt ($M = 11,78$, $SD = 1,68$; GB: $M = 13,47$, $SD = 1,63$; DE: $M = 10,87$, $SD = 0,78$). Da eine der untersuchten, britischen Schulen eine reine Mädchenschule war, bestand die britische Stichprobe zum Großteil aus Mädchen (72,03% weiblich, 24,44% männlich sowie 3,54% andere oder „möchte ich nicht sagen“). In der Stichprobe aus Deutschland waren 48,11% Mädchen und 46,48% Jungen (5,41% andere oder „möchte ich nicht sagen“). Insgesamt folgt daraus ein Gesamtanteil von 56,44% Mädchen und 38,80% Jungen (4,76% andere oder „möchte ich nicht sagen“). In Großbritannien wurden die Daten an verschiedenen privaten weiterführenden Schulen und in Deutschland an verschiedenen staatlichen weiterführenden Schulen (Gymnasien, Gesamtschulen und eine Realschule) erhoben. An den britischen Schulen wurde der sozioökonomische Status in Anlehnung an Rose (2005) sowie Rose und Pevalin (2001) durch den Bildungsgrad der Eltern und die berufliche Tätigkeit des*der Haupteinkommensbezieher*in der Familie erfasst. Der Großteil der Eltern hatte als höchsten Bildungsabschluss einen (Fach-)Hochschulabschluss oder die (Fach-)Hochschulreife und arbeitete in Führungs-, Verwaltungs- und Fachberufen. In Großbritannien gaben die meisten Schüler*innen als Muttersprache Englisch (89,07%), Chinesisch (3,86%) oder

Russisch (1,45%) an. In ihrer Freizeit oder in der Familie sprachen 13,83% eine weitere Sprache (außer Englisch). Als Staatsbürgerschaft wurden britisch (85,69%), chinesisch (3,86%) oder russisch (1,61%) genannt. In Deutschland wurde der sozioökonomische Status nicht erhoben. Hier war die Nationalität der Schüler*innen größtenteils deutsch (93,13%), gefolgt von türkisch (0,86%). Neben Deutsch (75,77%) waren die häufigsten Muttersprachen Türkisch (4,38%), Russisch (3,95%), Arabisch (1,89%), Polnisch (1,72%) und Serbisch (1,20%). Mehr als ein Viertel (27,66%) sprach außerhalb der Schule eine weitere Sprache neben Deutsch.

Die verwendeten Tests

Den in dieser Studie verwendeten Musiktests liegt die Annahme zugrunde, dass musikalische Fähigkeiten durch das Abschneiden in den Aufgaben, welche die verschiedenen Aspekte der Musikwahrnehmung testen, messbar sind. Musikalität als latente Variable bzw. Konstrukt ist nicht direkt beobachtbar und wird daher durch die Leistungen in den vier performativen Musiktests operationalisiert (vgl. [Werner et al., 2016](#), S. 949). Alle Tests basieren messtheoretisch auf der Item-Response-Theorie (IRT), die mit Ausnahme des Emotionsunterscheidungstests deren adaptive Verwendung ermöglicht. Das bedeutet, dass für jede Versuchsperson während der Testung ein eigenes Itemset dynamisch ausgewählt wird und sich die Schwierigkeit der Items somit individuell nach der Fähigkeit jeder einzelnen Person richtet ([Frey, 2012](#), S. 275). Die Tests nutzen außerdem realistische und damit ökologisch valide Stimuli aus dem den Schüler*innen vertrauten Rock-Pop-Bereich ([Müllensiefen & Hemming, 2018](#), S. 113). Für die Lösung der Testaufgaben werden nicht zwingend Fähigkeiten benötigt, die „an das Erlernen eines Instruments oder an formalen Musikunterricht gekoppelt sind“ ([Müllensiefen & Hemming, 2018](#), S. 93). Alle Hörtests – bis auf den Emotionsunterscheidungstest – beinhalteten Beispiele und eine Trainingsphase. [Tabelle A1](#) im Anhang gibt einen Überblick über die vier verschiedenen Tests, deren Reliabilitätswerte, die Größe der Stichprobe dieser Studie und Referenzen zur Kalibrierungsstudie.

Beat-Wahrnehmungstest (BAT)

Der *Computerised Adaptive Beat Alignment Test* (CA-BAT, kurz BAT) bewertet die Fähigkeit einer Person, den Grundschatz bzw. Beat in einem kurzen Musikausschnitt wahrzunehmen und danach zu beurteilen, ob dieser mit einem gleichzeitig erklingenden Metronom übereinstimmt. Beat-Wahrnehmung bezeichnet hierbei den Prozess, einen zugrunde liegenden Puls oder Grundschatz aus einem Musikausschnitt abzuleiten ([Harrison & Müllensiefen, 2018](#), S. 1; [Patel & Iversen, 2014](#), S. 2). In einer 2AFC-Aufgabe (Two-Alternative-Forced-Choice-Aufgabe) müssen die Versuchspersonen entscheiden, in welcher von zwei Versionen die Beep-Töne exakt mit dem Beat synchronisiert sind.

Emotionsunterscheidungstest (EDT)

Der Emotionsunterscheidungstest (engl. „Musical Emotion Discrimination Task“) bestimmt die Fähigkeit einer Person, den beabsichtigten emotionalen Ausdruck aus dargebotener Musik gemäß dem Production-Recognition-Paradigma zu erkennen ([MacGregor & Müllensiefen, 2019](#)). Dafür gibt die Person an, welcher von zwei Musikausschnitten eher eine bestimmte Emotion transportiert. Ein Item besteht hierbei aus zwei Versionen derselben Phrase, die sich nur in der intendierten Emotion unterscheiden. Die Aufgabe der Testpersonen ist es, in einem 2AFC-Format auszuwählen, welche der beiden Versionen eher mit einer der vier Zieleemotionen übereinstimmt, also beispielsweise, welche Version einen fröhlicheren Ausdruck hat.

Melodieunterscheidungstest (MDT)

Der Melodieunterscheidungstest (engl. „Melodic Discrimination Test“) erfasst die Fähigkeit einer Person, Unterschiede in verschiedenen kurzen Melodien zu erkennen (Harrison et al., 2017, S. 2), indem getestet wird, wie gut Veränderungen in einem Ton bei einer von insgesamt drei transponierten Versionen derselben Melodien identifiziert werden können (Harrison et al., 2017).

Intonationswahrnehmungstest (MPT)

Der Intonationswahrnehmungstest (engl. „Mistuning Perception Test“) misst die Fähigkeit, unterscheiden zu können, ob eine Gesangsstimme zu hoch oder zu tief im Verhältnis zur Hintergrundmusik intoniert ist. In einer 2AFC-Aufgabe muss entschieden werden, welche von zwei Versionen desselben Musikausschnitts falsch, also „verstimmt“ klingt (Larrouy-Maestri et al., 2019).

Goldsmiths Musical Sophistication Index (Gold-MSI)

Der *Goldsmiths Musical Sophistication Index* ist ein Selbstauskunftsfragebogen für Nicht-Musiker*innen, der musikalische Erfahrungheit nicht nur im Sinne musikalischer Expertise durch Instrumentalspiel und -unterricht misst, sondern auch als Folge weiterer allgemeiner, aktiver Auseinandersetzung und Umgangsweisen mit Musik in all ihren Facetten (Schaal et al., 2014, S. 423; Müllensiefen et al., 2014, S. 1). Das psychometrische Konstrukt des *Gold-MSI* gliedert sich in einen globalen Faktor (*GMSI*) und die fünf Faktoren *F1 Aktiver Umgang mit Musik*, *F2 Musikalische Wahrnehmungsfähigkeiten*, *F3 Musikalische Ausbildung*, *F4 Emotionen* und *F5 Gesangsfähigkeiten* (Müllensiefen et al., 2014; Müllensiefen & Hemming, 2018, S. 98; Übersetzungen aus Schaal et al., 2014).

Intelligenztest (MIQ)

Der Intelligenztest *MIQ* (engl.: „Matrix Reasoning IQ Test“; Chan & Kosinski, 2015) ist ein nonverbaler, computergestützter, adaptiver Test, der fluide Intelligenz, nonverbales Schlussfolgern und abstrakte Problemlösungsfähigkeit in neuen Situationen misst (Müllensiefen et al., 2015, S. 4). Ziel ist es, die deduktive Fähigkeit – aus etwas Abstraktem einen Sinn abzuleiten – zu erfassen (Raven, 2008, S. 22). Er basiert auf Ravens progressiven Matrizen (Raven, 2008; Chan & Kosinski, 2015). Die Items des *MIQ* zeigen eine Grafik, in der in einer 3×3-Matrix acht Elemente aus abstrakten einfachen Mustern dargestellt sind. Das neunte Element fehlt. Die Aufgabe besteht darin, aus acht möglichen anderen Mustern jenes auszuwählen, das die Matrix vervollständigt (Raven, 2008, S. 18; Müllensiefen et al., 2015, S. 4). In der *LongGold*-Studie wurden acht Matrizen (Items) benutzt.

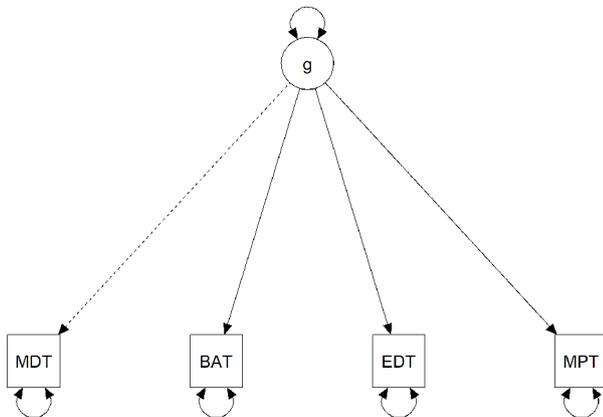
Strukturmodelle

Im Folgenden werden mögliche Strukturmodelle, die für die Beschreibung der Struktur musikalischer Fähigkeiten in Frage kommen und sich in der Anzahl der Faktoren unterscheiden, beschrieben. Die Ergebnisse der vier Musikalitätstests bilden hierbei die vier manifesten Variablen. Analog zu Spearman (1904b) Generalfaktormodell der Intelligenz und dem Generalfaktormodell der Musikalität, so wie es Wing (1961) seinem Musiktest zugrunde legte, ergibt sich als erstes potenzielles Modell, das 1×4-Modell, das schematisch in *Abbildung 1* dargestellt ist. Es enthält einen Musikalitätsfaktor, der die vier verbundenen manifesten Variablen – gemessen durch die vier Musikalitätstests – beeinflusst. Mit zwei Freiheitsgraden ist es überidentifiziert. Seashore (1919b) hingegen favorisierte eine multifaktorielle Theorie mit mehreren voneinander unabhängigen musikalischen Fähigkeiten. Davon und von Gardners (2006) Konzept der multiplen Intelligenzen sind die Modelle 2a und 2b

abgeleitet (siehe [Abbildung 2](#)): In diesen 4×1-Modellen bilden die vier Musikalitätstests vier einzelne Faktoren. In Modell 2a werden diese vier Faktoren als statistisch abhängig angenommen. Modell 2a und 1 sind hierarchisch geschichtet. Weil in Modell 2a mit abhängigen Faktoren die Anzahl der zu schätzenden Parameter genauso groß ist wie die Anzahl der empirischen Informationen ($df = 0$), ist es genau identifiziert ([Werner et al., 2016, S. 964](#)). Betrachtet man ein 4×1-Modellen mit statistisch unabhängigen Faktoren, erhält man das überidentifizierte Modell 2b.

Abbildung 1

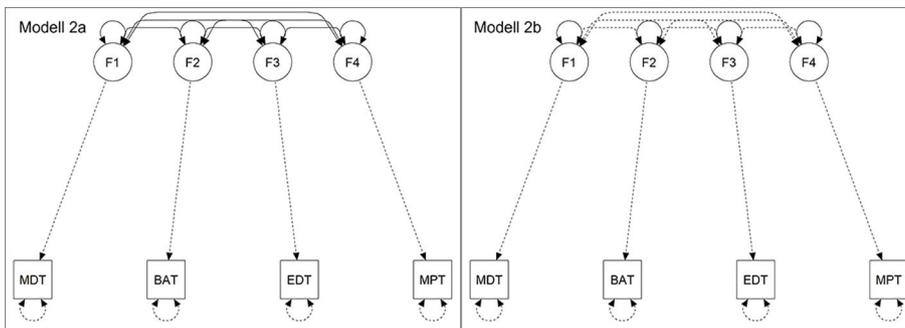
Das 1×4-Modell (Modell 1): Ein Faktor mit vier Indikatorvariablen



Anmerkung. Die Faktorladung der ersten Indikatorvariable (MDT) muss nicht geschätzt werden, da sie zum Festlegen der Maßeinheit der latenten Variable auf den Wert 1 gesetzt wurde. Aus diesem Grund ist sie mit einer gestrichelten Linie dargestellt.

Abbildung 2

Die 4×1-Modelle mit vier voneinander abhängigen bzw. unabhängige Faktoren mit je einer Indikatorvariable



Anmerkung. Die Faktorladung jeder latenten Variable wird jeweils auf den Wert 1 gesetzt, um Maßeinheiten für die latente Variable sinnvoll festlegen zu können ([Werner et al., 2016, S. 954](#)). Die Messfehlervarianzen jedes Faktors werden auf den Wert 0 gesetzt ([Beaujean, 2014, S. 40](#)). Da diese Werte nicht geschätzt werden müssen, sind sie mit gestrichelten Linien dargestellt. Dies gilt auch für die Korrelationen zwischen den unabhängigen Faktoren in Modell 2b.

Insgesamt gibt es weitere sieben theoretisch denkbare Modelle von Musikalität mit zwei latenten Faktoren und einer unterschiedlichen Verteilung der manifesten Variablen auf diese Faktoren (2 + 2 oder 3 + 1). Rein kombinatorisch wären noch weitere sechs Modelle mit insgesamt drei Faktoren (jeweils zwei Faktoren mit einem Indikator und ein Faktor mit zwei Indikatoren) denkbar. Diese werden jedoch nicht berücksichtigt, weil sie unteridentifiziert sind. Das rührt daher, dass es mehr zu schätzende Parameter (in diesem Fall elf: ein

Pfadkoeffizient, vier Messfehlervarianzen sowie die drei Varianzen und drei Kovarianzen der drei latenten Variablen) als nicht-redundante Informationen aus den Daten (in diesem Fall zehn: vier Varianzen und sechs Kovarianzen) gibt. Da die Intelligenzmodelle von Spearman (1904b) und Gardner (2006) sowie Wings (1961) und Seashores (1919b) Konstruktionen für Musikalitätstests ein einfaktorielles oder ein Modell mit vier Faktoren implizieren, bleiben diese weiteren Modelle hier unberücksichtigt und nur das 1×4-Modell und die 4×1-Modelle werden in Hinblick auf ihre Passung zu den Daten evaluiert.

In der Vergangenheit wurden musikalische Fähigkeiten oft durch Aufgaben gemessen, in denen Melodien voneinander unterschieden oder memoriert werden sollten (z. B. Michaelis, 1805, und Gordon, 1989). Daher misst der *MDT* im Vergleich zum *BAT*, *EDT* und *MPT* am ehesten das, was traditionellerweise in Musikalitätstests gemessen und unter dem Konstrukt *Musikalität* verstanden wurde. Aus diesem Grund wird bei allen Modellen die Testleistung des *MDTs* auf die erste manifeste Variable gesetzt (Werner et al., 2016, S. 955), die die Einheit der latenten Variable vorgibt und in *Abbildung 1* mit einer gestrichelten Linie erscheint.

Ergebnisse

Deskriptive Statistiken der Musikalitätstests

Tabelle 1 gibt die deskriptiven Statistiken der verwendeten Leistungstests sowie der Schulnoten und der selbst eingeschätzten musikalischen Expertise anhand des *Gold-MSI* wieder.

Tabelle 1

Deskriptive Statistiken der Musikalitäts- und Intelligenztests, der Schulnoten sowie des Globalfaktors (GMSI) und der Teilskalen F1 bis F5 des Gold-MSI

Variable	<i>n</i>	Fehlend	<i>M</i>	<i>SD</i>	Minimum	Maximum
Beat-Wahrnehmungstest	1786	0	-1,23	1,31	-4,00	2,15
Emotionsunterscheidungstest	699	1087	0,78	0,15	0,28	1,00
Melodieunterscheidungstest	1750	36	-1,23	1,14	-3,80	2,71
Intonationswahrnehmungstest	1786	0	-0,43	1,18	-4,00	2,66
Intelligenztest	1786	0	-1,74	1,08	-4,00	1,71
Schulnoten (in %)	1413	373	68,58	13,89	21,03	100,00
Allg. Musikal. Erfahrungheit (GMSI)	1694	92	3,81	0,87	1,00	6,72
<i>F1 Aktiver Umgang mit Musik</i>	1694	92	3,55	0,98	1,00	6,89
<i>F2 Musikalische Wahrnehmungsfähigkeiten</i>	1694	92	4,57	0,84	1,11	7,00
<i>F3 Musikalische Ausbildung</i>	1694	92	3,18	1,20	1,00	6,43
<i>F4 Emotionen</i>	1694	92	4,35	0,97	1,00	7,00
<i>F5 Gesangsfähigkeiten</i>	1694	92	4,22	1,03	1,00	7,00

Anmerkung. Da bei der Auswertung der Ergebnisse des *BATs*, *MDTs*, *MPTs* und *MIQs* die Item-Response-Theorie zum Tragen kam, liegen die Ergebnisse zwischen -4 und +4. Sie entsprechen dem geschätzten Personenparameter. Die Schulnoten sind auf Prozentwerte standardisiert. Die Items des Faktors *Musikalische Ausbildung* des *Gold-MSI* wurden auf einer Zustimmungsskala von 1 = „stimme ganz und gar nicht zu“ bis 7 = „stimme voll und ganz zu“ erhoben und gemittelt.

Korrelationen zwischen den Tests

In der Interkorrelationsmatrix in *Tabelle 2* lässt sich ablesen, wie stark je zwei der vier untersuchten Variablen miteinander korrelieren (unkorrigierte und korrigierte Pearson-Korrelationen unter- und oberhalb der Diagona-

le) und, wie reliabel die einzelnen Tests sind (Diagonale). Die Korrelationskoeffizienten zwischen den vier Variablen reichen von $r = ,34$ (Korrelation zwischen Beat-Wahrnehmungstest und Intonationswahrnehmungstest) bis $r = ,42$ (Melodieunterscheidungstest und Emotionsunterscheidungstest). Gemäß Ellis' (2010, S. 41) Benchmarks sprechen diese Ergebnisse für eine mittlere bis hohe Korrelation. Die anhand der Reliabilitäten der Kalibrierungsstudien korrigierten Korrelationen liegen zwischen $r = ,51$ und $r = ,65$. Watson (2017, S. 232) zufolge wird eine Interkorrelationsmatrix als für eine Faktorisierung geeignet erachtet, wenn die Mehrheit der Korrelationskoeffizienten zwischen ,20 und ,80 liegen. Dies ist der Fall.

Tabelle 2

Korrelationen zwischen den Musikalitätstests und deren Reliabilität

Test	1	2	3	4
1. Emotionsunterscheidungstest (EDT)	Rel. = ,690	,585	,646	,558
2. Melodieunterscheidungstest (MDT)	,422	Rel. = ,754	,536	,510
3. Intonationswahrnehmungstest (MPT)	,414	,359	Rel. = ,595	,564
4. Beat-Wahrnehmungstest (BAT)	,362	,346	,340	Rel. = ,610

Anmerkung. Unterhalb der Diagonalen stehen die Pearson-Korrelationen – berechnet aus je einem kompletten Wertepaar – zwischen den vier Testscores, die mit $p < ,001$ signifikant sind. Auf der Diagonalen stehen die Reliabilitäten (Rel.) der einzelnen Tests. Oberhalb der Diagonalen stehen die anhand der Reliabilitäten der Kalibrierungsstudien (Referenz zu den Studien siehe letzte Spalte in Tabelle A1) korrigierten Korrelationen (Minderungskorrektur; Spearman, 1904a, S. 90).

Modellierung der latenten Variablenmodelle

Die Hypothesen aus der Literatur über ein einfaktorielles Modell oder ein Modell mit vier unverbundenen Faktoren motivieren einen Modellvergleich zwischen dem 1×4-Modell und den 4×1-Modellen. Weil gegebenenfalls Unterschiede bestehen könnten, erfolgt der Vergleich getrennt für beide Länder. Daher sei zunächst ausschließlich die Stichprobe aus Großbritannien betrachtet. Eine explorative Faktorenanalyse der Daten der britischen Stichprobe legt Modell 1 nahe (Ergebnisse siehe Anhang). Die folgende Betrachtung untersucht Modell 1, Modell 2a und 2b in einer konfirmatorischen Faktorenanalyse auf Grundlage der Daten der britischen Stichprobe mit $n = 622$ für die vier Musikalitätstests *BAT*, *EDT*, *MDT* und *MPT* mit Full-Information-Maximum-Likelihood-Verfahren (Beaujean, 2014, S. 119 und Rosseel, 2012, S. 30) und robuster Maximum-Likelihood-Schätzmethode (Werner, 2015, S. 9) und überprüft das Ergebnis der explorativen Faktorenanalyse. Die latenten Variablen wurden standardisiert.

Ein Chi-Quadrat-Test untersucht „globale [...] Abweichungen zwischen Modell und Daten“ (Werner et al., 2016, S. 957) und beurteilt damit deren Passung. Das Ergebnis des Chi-Quadrat-Tests für das 1×4-Modell mit einem Musikalitätsfaktor und vier verbundenen manifesten Variablen (Modell 1) ist nicht signifikant (siehe Tabelle 3 zweite und dritte Spalte, $p = ,391$). Das bedeutet, dass die Nullhypothese beibehalten wird (Werner et al., 2016, S. 967), die Annahmen des Modells richtig sind (vgl. Werner et al., 2016, S. 958) und Modell 1 und Daten nicht signifikant voneinander abweichen (Werner et al., 2016, S. 959). Alle weiteren Gütekriterien (*RMSEA*, *SRMR*, *GFI* und *CFI*) befinden sich innerhalb der jeweils passenden Wertebereiche der Schwellenwerte (siehe Tabelle 3 letzte Zeile). Zusammen deutet dies auf einen sehr guten Modellfit für das 1×4-Modell hin.

Tabelle 3

Robuste Güte- und Informationskriterien der Modelle für die Maximum-Likelihood-Schätzung mit robuster Standardfehlerschätzung (MLR)

Modell	χ^2			RMSEA			SRMR	BIC
	Wert	<i>p</i>	<i>df</i>	Wert	90% CI			
				<i>LL</i>	<i>UL</i>			
GB (n = 622)								
1 (1×4)	1,9	,391	2	,000	,000	,079	0,01	4552
2a (4×1; abhängig)	—	—	0	—	—	—	—	4563
2b (4×1; unabhängig)	309,7	< ,001	6	,296	,268	,324	0,23	4857
DE (n = 1164)								
1 (1×4)	0,4	,824	2	,000	,000	,036	0,01	10881
GB & DE (n = 1786)								
1 (1×4)	0,6	,725	2	,000	,000	,036	0,01	15777
Schwellenwerte								
	5,99 (<i>df</i> = 2) bzw. 12,59 (<i>df</i> = 6)	> ,05	—	≤ ,06	—	—	≤ 0,08	—

Anmerkung. χ^2 = Model Fit Test Statistic; SRMR = Standardized Root Mean Square Residual; BIC = Bayes'sches Informationskriterium. Goodness-of-Fit-Index und Comparative-Fit-Index von Modell 1 sind für alle Stichproben 1,000, für Modell 2a beide 1,00 und für Modell 2b 0,98 und 0,00, wobei der Schwellenwert bei $\geq 0,95$ liegt (GFI: Shevlin und Miles, 1998; CFI: Hu und Bentler, 1999). In der letzten Zeile sind die üblichen Schwellenwerte für gute Modellpassung angegeben (Hu und Bentler, 1999).

Da Modell 2a keine Freiheitsgrade hat ($df = 0$), kann kein Chi-Quadrat-Test durchgeführt werden. Stattdessen wird als Nächstes durch einen χ^2 -Differenztest die Frage geklärt, ob Modell 1 oder 2a besser zu den Daten passt. Ein Vergleich ist hier möglich, weil Modell 1 und 2a geschachtelt sind. Modell 1 wird gegen das Referenzmodell Modell 2a getestet. Das Ergebnis ist in Tabelle 3 (fünfte Zeile) abzulesen und zeigt, dass der Unterschied zwischen Modell 1 und 2a ($\Delta\chi^2 = 1,9$) nicht signifikant ist ($p = ,391$). Das bedeutet, dass Modell 1 keinen signifikant schlechteren Chi-Quadrat-Wert hat als das komplexere, abhängige 4×1-Modell. Da es gleichzeitig einen besseren BIC (Bayes'sches Informationskriterium) hat und weniger komplex ist als Modell 2a, wird das einfaktorielle Modell 1 dem multifaktoriellen Modell mit vier abhängigen Faktoren (Modell 2a) vorgezogen.

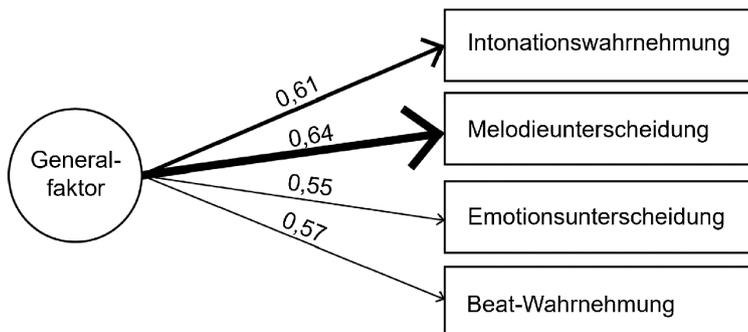
Im Folgenden wird Modell 2b bestehend aus vier unabhängigen Faktoren mit jeweils einer Indikatorvariable untersucht. Hier ist der Chi-Quadrat-Test signifikant, was auf eine Abweichung zwischen Modell und Daten schließen lässt (vgl. Werner et al., 2016, S. 958). Die restlichen Gütekriterien mit Ausnahme des Goodness-of-Fit-Index liegen jenseits der empfohlenen Bereiche. Das Bayes'sche Informationskriterium (BIC) ist für das einfaktorielle Modell 1 kleiner als für die multifaktoriellen Modelle (siehe Tabelle 3 letzte Spalte). Aus diesen Gründen und der Tatsache, dass die Leistungen in den Musikalitätstests mäßig bis stark miteinander korrelieren, können die 4×1-Modelle und die Möglichkeit, dass die Testleistungen als vier Faktoren unverbunden sind, ausgeschlossen werden.

Angesichts der Gütekriterien kann davon ausgegangen werden, dass Modell 1 korrekt spezifiziert ist und damit die Ergebnisse der Parameterschätzung sinnvoll inhaltlich interpretiert werden können (vgl. Werner et al., 2016, S. 958): Diese zeigen standardisierte Ladungen zwischen 0,55 und 0,64 und liegen somit in einem Bereich,

der in der Literatur generell akzeptiert wird (vgl. hierzu die Faktorladungen bei Wing, 1941, Tabelle 2, S. 34 und bei Carroll, 1993, Tabellen 9.5 und 9.6). Die Variablen *MDT* und *MPT* haben die höchsten Ladungen (,64 und ,61). Das heißt, dass sie am stärksten mit dem Konstrukt *Musikalität* zusammenhängen und es am besten abbilden (vgl. Werner et al., 2016, S. 951). Die Kommunalitäten der Faktorladungen sind niedrig (,30 bis ,41) bzw. die Uniqueness (standardisierte Messfehlervarianzanteile der einzelnen Tests) hoch (,59 bis ,70), was bedeutet, dass das latente Konstrukt *Musikalität* die Zusammenhänge zwischen den vier Musikalitätstests nicht vollständig erklärt. Da die beiden 4×1-Modelle verworfen wurden und das 1×4-Modell mit einem Faktor (siehe [Abbildung 3](#)) einen guten Modellfit hat, wird es im Anschluss favorisiert.

Abbildung 3

Faktorstruktur als Ergebnis der konfirmatorischen Faktorenanalyse mit standardisierten Ladungen (britische Stichprobe)



Anmerkung. Je breiter die Pfeile sind, desto größer ist die Faktorladung.

Insgesamt kann die Schlussfolgerung gezogen werden, dass eine Operationalisierung des Konstrukts *Musikalität* durch nur einen Generalfaktor – bestehend aus den Leistungen der vier Tests *BAT*, *EDT*, *MDT* und *MPT* – möglich ist und das Konstrukt *Musikalität* als latente Variable des einfaktoriellen Strukturgleichungsmodells aufgefasst werden kann. Das Generalfaktormodell passt gut auf die Daten, wobei der Generalfaktor die Zusammenhänge zwischen den vier Musikalitätstests nicht vollständig erklärt. Dies kommt daher, dass alle Tests verschiedene Aspekte musikalischer Wahrnehmungsfähigkeiten messen. Die Residualvarianzen der einzelnen Tests sind demnach unter anderem auf deren spezielle Zielkonstrukte zurückzuführen.

Übertragung auf die Stichprobe aus Deutschland

Im nächsten Schritt wird überprüft, ob das gefundene Ergebnis der Generalfaktorhypothese auf die in Deutschland erhobene Stichprobe übertragen werden kann und ob damit eine Generalisierung möglich ist. Eine konfirmatorische Faktorenanalyse überprüft, ob die Korrelationen zwischen den Ergebnissen der vier Musikalitätstests aus dem Einfluss eines zugrunde liegenden Generalfaktors resultieren wie im Falle der britischen Stichprobe. Dafür wird analog zur vorherigen Analyse Modell 1 (ein Faktor mit den vier Indikatorvariablen *BAT*, *EDT*, *MDT* und *MPT*) analysiert.

In [Tabelle 3](#) (neunte Zeile) sind die Güte- bzw. Informationskriterien der Modellierung des latenten Variablenmodells anhand der deutschsprachigen Stichprobe berichtet. Die Tabelle zeigt, dass Modell 1 gut zu den Daten passt, da der Chi-Quadrat-Test nicht signifikant ist und alle anderen Gütekriterien (*RMSEA*, *SRMR*, *GFI* und *CFI*) ober- oder unterhalb des jeweiligen Schwellenwerts liegen. Die standardisierten Ladungen der

Indikatorvariablen auf dem Generalfaktor g liegen zwischen ,49 und ,56. Die Variablen *EDT* und *MDT* sind am engsten mit der latenten Variable g verknüpft und zeigten auch in der britischen Stichprobe den stärksten Zusammenhang mit g . Die Residualvarianzen der einzelnen Tests sind relativ hoch (zwischen ,68 und ,76), was auf Messfehlereinflüsse schließen lässt. Die Kommunalitäten der manifesten Variablen – also der Anteil ihrer Varianz, der durch die latenten Variablen des Modells erklärt wird – sind teilweise niedriger als in der vorherigen Analyse. Das bedeutet, dass das latente Konstrukt *Musikalität* die Varianzen der Indikatorvariablen nicht vollständig erklärt und der Einfluss des gefundenen Generalfaktors nicht alle Zusammenhänge zwischen den vier Musikalitätstests vollständig begründen kann.

Der musikalische Generalfaktor

Aus den bisherigen Ergebnissen dieser Studie geht folgende Überlegung hervor: Bei Annahme eines Zusammenhangs zwischen den Leistungen in den unterschiedlichen Tests ist dieser durch den Einfluss eines und nicht mehrerer zugrunde liegender Faktoren begründet. Damit sprechen die Ergebnisse der Analysen beider Stichproben für die Generalfaktorhypothese.

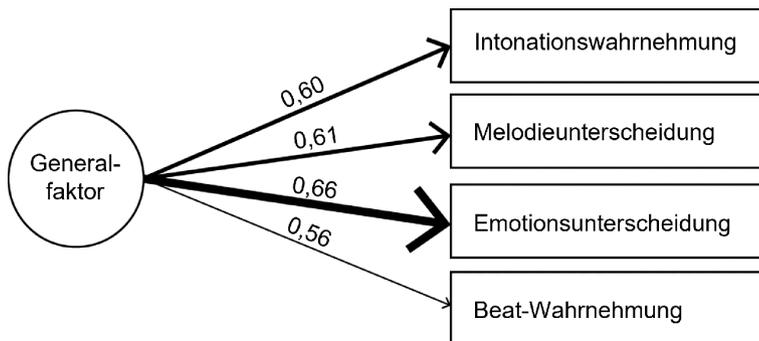
Bildung des Generalfaktors

Der im ersten und zweiten Teil der Studie identifizierte Generalfaktor – bestehend aus den vier Variablen *BAT*, *EDT*, *MDT* sowie *MPT* – diente anschließend als neue Variable für weitere Analysen. Daher wurde eine dritte konfirmatorische Faktorenanalyse eines einfaktoriellen, latenten Variablenmodells (entsprechend Modell 1, siehe [Abbildung 1](#)) auf Grundlage der deutschen und britischen Stichprobe durchgeführt und die Faktorwerte der latenten Variable g als neue Variable *GF* durch die Regressionsmethode extrahiert. Dank der Verwendung des Full-Information-Maximum-Likelihood-Verfahrens war dies auch für Fälle mit fehlenden Werten möglich. Die Güte- und Informationskriterien der Modellierung stehen in [Tabelle 3](#) (elfte Zeile) und sprechen für eine gute Modellpassung. Die Ergebnisse der Parameterschätzung für das Generalfaktormodell sind in [Abbildung 4](#) abgedruckt. Die standardisierten Ladungen der vier Indikatorvariablen liegen im Bereich zwischen ,56 und ,66, der in der Literatur generell akzeptiert wird (vgl. hierzu die Faktorladungen bei [Wing, 1941](#), Tabelle 2, S. 347 und bei [Carroll, 1993](#), Tabellen 9.5 und 9.6). Die höchste Ladung hat die Variable *EDT* mit ,66. Das entspricht dem Ergebnis der Faktorenanalyse der deutschen Stichprobe, in der die Variable *EDT* ebenfalls die höchste Ladung (,56) hatte.

Die Werte der Kommunalität bewegen sich zwischen ,32 und ,44 und sind damit höher als in den vorherigen Faktorenanalysen. Das bedeutet, dass durch die latente Variable g , also durch den Generalfaktor, insgesamt teilweise mehr Varianz der manifesten Variablen erklärt werden konnte als in den separaten Faktorenanalysen der britischen und der deutschen Stichproben. Trotzdem liegt der Anteil an durch g unerklärter Varianz bei allen Indikatorvariablen über 50%. Die geschätzten Faktorwerte der latenten Variable g , die als Generalfaktor aufgefasst wird, haben den Mittelwert 0, da sie standardisiert sind ($SD = 0,558$). Die Faktorwerte liegen zwischen $-1,706$ und $1,527$. Die Verteilung der Variable *GF* weicht von einer Normalverteilung ab (vgl. [Abbildung 5](#) und Shapiro-Wilk-Test: $W = 0,99$, $p < ,001$).

Abbildung 4

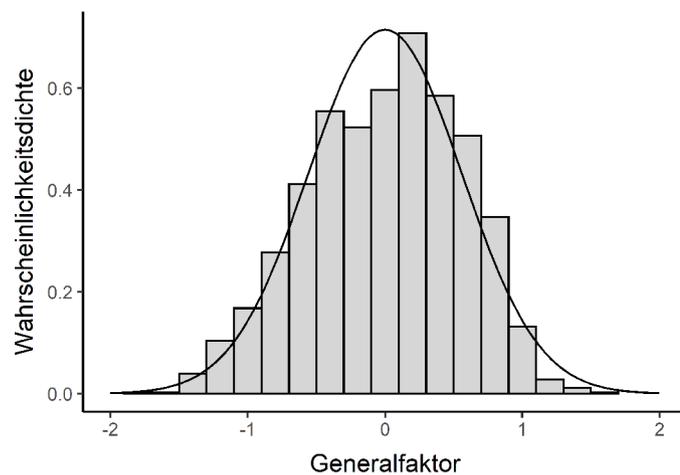
Faktorstruktur als Ergebnis der konfirmatorischen Faktorenanalyse mit standardisierten Ladungen (britische und deutsche Stichprobe)



Anmerkung. Je breiter die Pfeile, desto größer ist die Faktorladung.

Abbildung 5

Histogramm mit Dichtefunktion der Normalverteilung



Als Nächstes kann nun der Zusammenhang zwischen dem ebenfalls für die Gesamtstichprobe identifizierten Generalfaktor und anderen interessierenden Merkmalen untersucht werden und es können Normen für die Musikalitätstests generiert werden.

Korrelation des Generalfaktors mit anderen Variablen

Tabelle 4 informiert über die Korrelationen zwischen dem Generalfaktor, der Leistung im Intelligenztest *MIQ*, den Schulnoten, dem Globalfaktor der *Allgemeinen Musikalischen Erfahrung* des *Gold-MSI* (GMSI; Müllensiefen et al., 2014), den Faktoren *Aktiver Umgang mit Musik*, *Musikalische Wahrnehmungsfähigkeiten*, *Musikalische Ausbildung*, *Emotionen* und *Gesangsfähigkeiten* des *Gold-MSI*.

Tabelle 4

Interkorrelationen (Pearson) des Generalfaktors (GF) mit dem Intelligenztest MIQ, den Schulnoten, der Globalskala Allgemeine Musikalische Erfahrungheit (GMSI) und den Gold-MSI-Teilskalen (F1 bis F5)

Tests und Faktoren	1	2	3	4	5	6	7	8
1. Generalfaktor	—							
2. Intelligenztest MIQ	,396*	—						
3. Schulnoten	,168*	,150*	—					
4. Allg. Musikal. Erfahrungheit (GMSI)	,286*	,059*	,127*	—				
5. F1 Aktiver Umgang mit Musik	,202*	,037	,025	,769*	—			
6. F2 Musikalische Wahrnehmungsfähigkeiten	,323*	,111*	,162*	,676*	,478*	—		
7. F3 Musikalische Ausbildung	,291*	,087*	,139*	,757*	,488*	,432*	—	
8. F4 Emotionen	,296*	,129*	,096*	,579*	,585*	,515*	,344*	—
9. F5 Gesangsfähigkeiten	,190*	,007	,081*	,804*	,565*	,589*	,397*	,456*

Anmerkung. $1330 \leq df \leq 1780$.

* $p < ,05$.

Der extrahierte Generalfaktor korreliert mäßig mit der Teilskala *Musikalische Wahrnehmungsfähigkeiten* ($r = ,32$, $p < ,001$; vgl. Müllensiefen et al., 2014). Dies spiegelt die Tatsache wider, dass alle hier verwendeten Musikalitätstests Aspekte der Musikwahrnehmung untersuchen. Die mäßig starke Korrelation mit der Teilskala *Emotionen* des *Gold-MSI* ($r = ,30$, $p < ,001$) stützt das Ergebnis der Faktorenanalyse mit der deutschen und der gesamten Stichprobe, wonach die Ergebnisse des *EDTs* die latente Variable *Musikalität g* am besten abbilden. Außerdem korreliert der Generalfaktor *GF* mäßig stark mit dem Globalfaktor des *Gold-MSI* und dessen Teilskala *Musikalische Ausbildung* (beide $r = ,29$, $p < ,001$) sowie mäßig bis stark mit dem Intelligenztest *MIQ* ($r = ,40$, $p < ,001$).

Intelligenz und *Musikalische Ausbildung* (Faktor *F3*) korrelieren nur gering miteinander ($r = ,09$, $p < ,001$). Diese Korrelation verschwindet und ist nicht mehr signifikant, wenn man den Generalfaktor kontrolliert und die partielle Korrelation zwischen *MIQ* und *F3* betrachtet ($r_{(MIQ, F3 | GF)} = -,03$, n. s.). Dies bedeutet, dass entweder die musikalische Ausbildung die Musikalität beeinflusst und diese wiederum Einfluss auf die Intelligenz hat oder das umgekehrt die Intelligenz die Musikalität beeinflusst und diese dann wiederum Auswirkungen darauf hat, wie viel musikalische Ausbildung eine Person erhält. Wir wissen also nicht, ob hohe kognitive Fähigkeiten zu mehr musikalischer Ausbildung führen oder umgekehrt.

Normwerte der Referenzstichprobe für BAT, MDT und MPT

Die Normierung eines Tests dient der Erstellung eines objektiven „Bezugssystems, mit dessen Hilfe die Ergebnisse einer Testperson im Vergleich zu den Merkmalsausprägungen anderer Personen eindeutig eingeordnet und interpretiert werden können“ (Moosbrugger & Kelava, 2012, S. 19). Im vorliegenden Fall besteht die Zielpopulation aus Jugendlichen der Sekundarstufe I in Deutschland und Großbritannien, die einen ähnlichen sozioökonomischen Hintergrund wie die hier untersuchte und im Methodenteil beschriebene Stichprobe aufweisen.

Es wird angenommen, dass musikalische Entwicklung im Jugendalter – und damit die Entwicklung der getesteten Fähigkeiten – durch Alter, Geschlecht und musikalische Ausbildung beeinflusst wird (Hargreaves & Lamont, 2017; vgl. Abbildung 6). Um diese Vermutung empirisch zu untersuchen, wird musikalische Entwicklung durch

das Modell einer nicht-parametrischen Regression beschrieben, wobei der jeweilige Testwert (z. B. *BAT*) die abhängige Variable und die Variablen *Alter*, *Geschlecht* und *Musikalische Ausbildung* (Faktor *F3* des *Gold-MSI*) die unabhängigen Variablen bilden. Die Prüfung der Linearität zwischen den abhängigen Variablen und den unabhängigen Variablen *Alter* und *Musikalische Ausbildung* erfolgt graphisch in Lowess-Diagrammen (vgl. [Abbildung 6](#)). Da keine lineare Beziehung zwischen den Variablen herrscht, wird ein non-parametrisches Modell verwendet. Die Annahme des Modells ist hierbei, dass die Variable des jeweiligen Testscores abhängig von den Variablen *Alter*, *Geschlecht* und *Musikalische Ausbildung* ist. Dies wurde in einem Unabhängigkeitstest mit der Maximum-Statistik als Teststatistik untersucht (R-Paket *coin*, Funktion *independence_test*). Dessen Nullhypothese ist, dass zwei Gruppen von Variablen y_1, \dots, y_q und x_1, \dots, x_p , die auf beliebigen Skalen gemessen wurden, voneinander unabhängig sind ([Hothorn et al., 2021, S. 25](#)). Das Modell einer nicht-parametrischen Regression und ein Unabhängigkeitstest wurden einer multiplen linearen Regression oder einer Korrelation vorgezogen, da der Anstieg in der abhängigen Variable und in den Variablen *Alter*, *Geschlecht* und *Musikalische Ausbildung* nicht linear ist. Das Ergebnis zeigt, dass Unterschiede in allen unabhängigen Variablen zu Unterschieden in der abhängigen Variable (Testscores) führen ($p < ,001$). Man kann daher davon ausgehen, dass die Testleistungen in den Musikalitätstests tatsächlich mit dem Alter, dem Geschlecht und dem Ausmaß an musikalischer Ausbildung zusammenhängen. Aus diesem Grund wurden für alle Kombinationen der unabhängigen Variablen *Alter*, *Geschlecht* und *Musikalische Ausbildung* die Normen der Tests *BAT*, *MDT* und *MPT* entsprechend differenziert, wobei die ursprünglich metrische Variable *Musikalische Ausbildung* in eine ordinale Variable umgewandelt wurde.

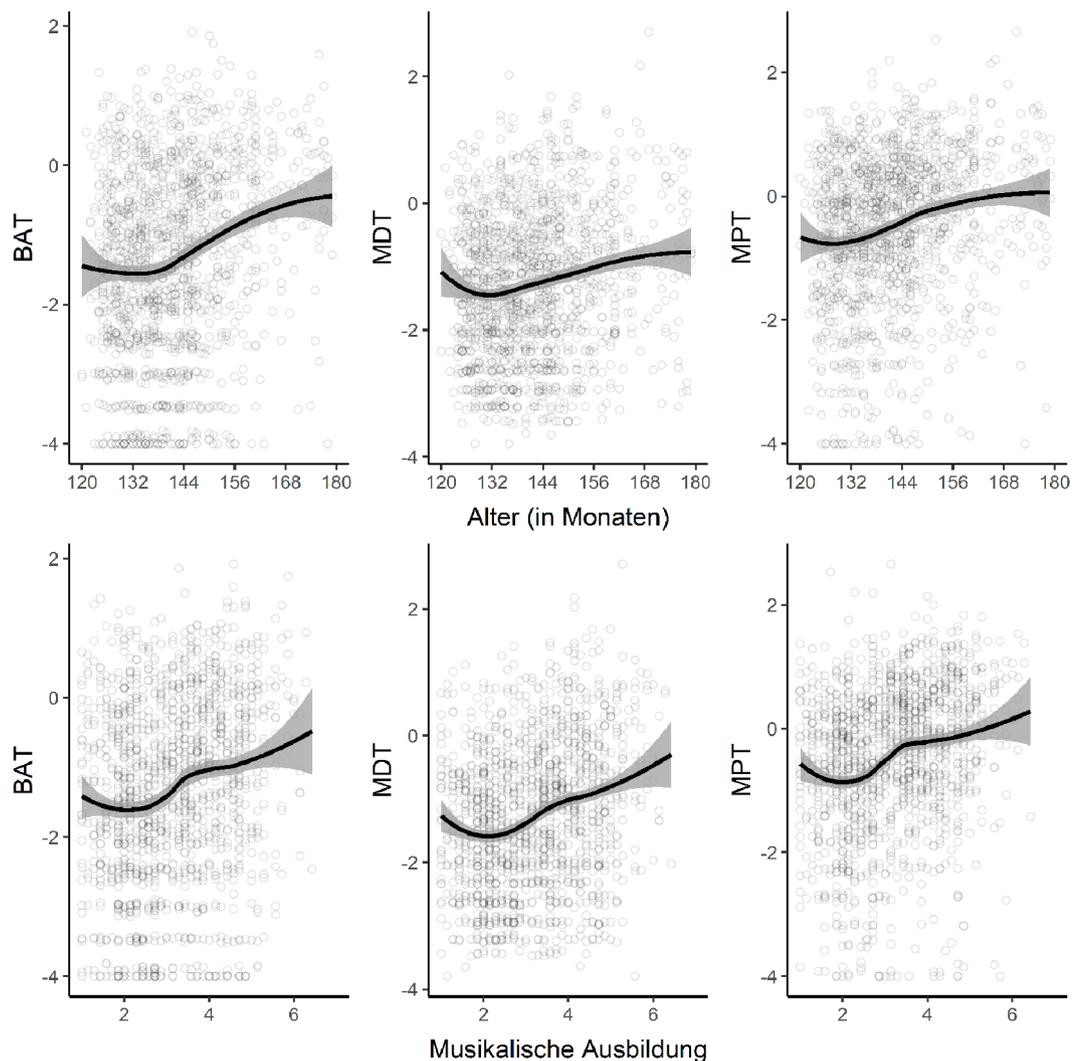
Für die Auswahl der Referenzstichprobe des *BATs*, des *MDTs* und des *MPTs* wurden aus der deutschen und britischen Stichprobe (zusammen $N = 1786$) alle Fälle ausgeschlossen, die fehlende Werte für den Faktor *Musikalische Ausbildung* des *Gold-MSI* ($n = 92$) oder für die Variable *Alter* ($n = 1$) hatten oder die als Geschlecht nicht männlich oder weiblich angegeben hatten ($n = 85$), weil später Normen unter anderem nach Geschlechtern getrennt erstellt wurden. Diese zusätzlich nach Alter und musikalischer Ausbildung aufzuteilen, hätte sehr kleine Substichproben mit $n < 50$ ergeben. Die verbliebenen 1614 Fälle wurden als Nächstes in Subgruppen nach Geschlecht (weiblich und männlich) und Alter eingeteilt (z. B. alle zehnjährigen Mädchen; vgl. [Tabelle A2](#)). Daraufhin wurden alle Subgruppen ausgeschlossen, die kleiner als $n = 50$ waren. Dieser Grenzwert wurde gewählt, um sicherzustellen, dass eine Substichprobe aus mindestens zwei Schulklassen mit $n \approx 25$ besteht, und um so einen Klassenbias zu vermeiden. Das Ergebnis war eine Referenzstichprobe aus $n = 1443$ Fällen unterteilt in neun Subgruppen (für die Einteilung der Subgruppen und deren Größen siehe Spalte eins bis drei in [Tabelle A2](#)). Waren diese Subgruppen größer oder gleich $n_{\text{Subgruppe}} = 100$, wurden sie anschließend in mehrere Substichproben anhand der Variable *Musikalische Ausbildung* geteilt, sodass jede Substichprobe mindestens 50 Fälle umfasste. Im Folgenden wird durch die Begriffe „Subgruppe“ und „Substichprobe“ zwischen den nur nach Alter und Geschlecht (Subgruppe) und den zusätzlich nach *Musikalischer Ausbildung* eingeteilten Stichproben (Substichproben) unterschieden. Wenn die Subgruppe größer oder gleich $n = 100$ ist, bilden mehrere Substichproben eine Subgruppe. Zur Festlegung eines Kriteriums für die Einteilung in k ungefähr gleich große, endgültige Substichproben, wurden die Ergebnisse der Variable *Musikalische Ausbildung* in k Leistungsstufen bzw. Level sortiert und so Quantile gebildet. Die Variablen, für welche die Normen erstellt wurden, sind in den jeweiligen Substichproben nicht normalverteilt.

Die Normwerte sind entsprechend der Subgruppen in [Tabelle A2](#) in Form von Mittelwerten und Standardabweichungen aufgelistet. Sie erlauben eine näherungsweise Einschätzung des Leistungsniveaus eines Jugendlichen in Abhängigkeit zu Alter, Geschlecht und musikalischer Ausbildung. Wenn beispielsweise ein zehnjähriger

Junge auf der Teilskala *Musikalische Ausbildung* zwei Punkte (Level 1) und im *BAT* einen Testscore von $-0,70$ erreicht, bedeutet dies, dass er, dafür, dass er keinerlei musikalische Ausbildung genießt, eine überdurchschnittliche Beat-Wahrnehmungsfähigkeit im Vergleich zur Referenzgruppe besitzt. Denn für zehnjährige Jungen ohne musikalisches Training (Level 1) ist der Normmittelwert $-1,82$ (Standardabweichung = $1,08$) und der Wert $-0,70$ somit mehr als eine Standardabweichung vom Mittelwert entfernt. Die Normen für den *EDT* werden in dieser Arbeit nicht berichtet, da der *EDT* zum Zeitpunkt der Datenauswertung weiterentwickelt wurde (MacGregor, Ruth & Müllensiefen, 2021), sodass damit zu rechnen ist, dass in Zukunft ein neuer (adaptiver) Emotionsunterscheidungstest mit neuen Normwerten verfügbar sein wird.

Abbildung 6

Loess-Diagramme der Testscores im Verhältnis zum Alter (in Monaten) und zur musikalischen Ausbildung (Faktor F3 des Gold-MSI)



Anmerkung. Loess-Anpassungslinien beruhen auf einer non-parametrischen, lokal gewichteten Regressionsmethode mit smoothing-Eigenschaften.

Diskussion

Das wesentliche Ziel der vorliegenden Untersuchung war, eine Antwort auf die Frage nach der Struktur von Musikfähigkeit zu finden. Der Stand der Forschung zu diesem Thema ist noch nicht zufriedenstellend, da es konkurrierende ein- und multifaktorielle Modelle (vgl. Wing, 1961, vs. Seashore, 1919b) gibt. Im ersten Teil dieser Studie wurden die Daten einer großen englischsprachigen Stichprobe für vier Musikalitätstests zur Beat- und Intonationswahrnehmung sowie zur Melodie- und Emotionsunterscheidung analysiert, deren Ergebnisse mäßig stark miteinander korrelieren. In Analogie zu ein- und multifaktoriellen Modellen der Intelligenz wurden mehrere Modelle aufgestellt, die diese Korrelation als Folge des Einflusses eines oder mehrerer dahinterliegender Faktoren und im Zuge dessen die Struktur musikalischer Fähigkeiten beschreiben. Aufgrund der Ergebnisse eines χ^2 -Differenztests und eines Chi-Quadrat-Tests und der Korrelationen zwischen den Leistungen in den vier Musiktests kann ein multiples Konstrukt aus verbundenen und unverbundenen musikalischen Einzelfähigkeiten – wie von Seashore angenommen – abgelehnt werden. Vielmehr spricht die Struktur der Korrelationen für ein Generalfaktormodell, welches im zweiten Teil dieser Studie anhand einer größeren, deutschsprachigen Stichprobe bestätigt werden konnte. Dies erlaubte es, im dritten Teil einen musikalischen Generalfaktor – bestehend aus den Leistungen der vier Musikalitätstests – zu formulieren. Die empirische Evidenz dieser Studie spricht demnach dagegen, dass musikalische Fähigkeiten modular organisiert sind, wie es Peretz und Coltheart (2003) für die Verarbeitung von Musik annehmen, und trägt durch die Favorisierung des einfaktoriellen Modells zu einer klareren Theorielage bei. Anhand der Ergebnisse der ersten beiden Teile der Studie wurden die Faktorwerte eines Generalfaktors für die zusammengeführte Stichprobe berechnet und schließlich für die einzelnen Tests alters-, geschlechts- und übungsabhängige Normwerte ermittelt.

Die gemeinsame Varianz der Leistungen auf den vier Musikalitätstests konnte nur teilweise auf das latente Konstrukt *Musikalität* zurückgeführt werden. Gründe dafür könnten sein, dass sich die einzelnen Tests auf jeweils unterschiedliche Aspekte der Musikwahrnehmung beschränken und damit jeweils vermeintlich verschiedene Aspekte musikalischer Wahrnehmungsfähigkeit messen. Das liegt daran, dass sie unabhängig voneinander und nicht mit der Absicht konstruiert wurden, ein gemeinsames Konstrukt (allgemeine Musikalität) zu erfassen, sondern immer nur eine ganz bestimmte Teilfähigkeit (Referenzen zu den Studien siehe letzte Spalte in Tabelle A1). Dass die einzelnen Tests dennoch so hoch korrelieren und viel gemeinsame Varianz haben, weist nicht auf mehrere voneinander unabhängige Teilfähigkeiten hin, sondern darauf, dass diese Korrelationen durch nur eine dahinterliegende latente Fähigkeit, wie zum Beispiel einen musikalischen Generalfaktor, generiert werden könnten, welcher die Leistung auf allen Tests beeinflusst. Weitere Gründe dafür, dass der Generalfaktor die Zusammenhänge zwischen den vier Musikalitätstests nicht vollständig erklärt, könnten Mediatorvariablen wie Gedächtnis oder allgemeines Hörvermögen sein, die einen Einfluss auf die Leistungen in den Musikalitätstests haben. Schüler*innen mit hohen Werten im Generalfaktor sind tendenziell auch musikalisch ausgebildet, spielen also ein Instrument oder haben Instrumentalunterricht, weisen eine hohe allgemeine musikalische Erfahrung auf und schnitten im Intelligenztest gut ab. Anhand dieser Ergebnisse können allerdings keine Aussagen darüber getroffen werden, ob musikalische Ausbildung zu hohen Werten im Generalfaktor führt oder umgekehrt. Genauso wenig kann darüber spekuliert werden, ob intelligentere Schüler*innen wegen ihrer Intelligenz so hohe Generalfaktorwerte haben oder, ob ihre allgemeine Musikalität beeinflusst, wie gut sie im Intelligenztest abgeschnitten haben. Die Korrelation zwischen Generalfaktor und Leistung im Intelligenztest könnte dafür sprechen, dass darüber hinaus allgemeine Intelligenz eine Mediatorvariable darstellt. Das würde

bedeuten, dass intelligentere Schüler*innen möglicherweise die Aufgabenstellung der Musiktests besser verstehen, dadurch besser darin abschneiden und dementsprechend auch höhere Werte im Generalfaktor zeigen.

Alternativ kann man in der Korrelation auch eine Bestätigung der Zwei-Faktoren-Theorie der Intelligenz von Spearman sehen, nach der allgemeine Intelligenz und verschiedene Einzelfähigkeiten in Intelligenztests zusammenhängen (Neyer & Asendorpf, 2018, S. 155). Wenn man den Begriff der Einzelfähigkeiten auf die durch Musikalitätstests gemessenen Fähigkeiten ausweitet, ließe sich eine Korrelation zwischen allgemeiner Intelligenz und musikalischer Fähigkeit durch die Zwei-Faktoren-Theorie begründen. Eine weitere Erklärung wäre ein hierarchisches Intelligenzmodell, wie beispielweise die Cattell-Horn-Carroll-Theorie der kognitiven Fähigkeiten (CHC-Theorie; McGrew, 2005). Darin ist Musikalität ein Unterfaktor der auditorischen Fähigkeiten, die wiederum mit der allgemeinen Intelligenz verbunden sind.

Ausgangspunkt für diese Studie waren verschiedene Intelligenzmodelle und die Frage, ob Musikalität durch ein vergleichbares Modell beschrieben werden kann. Die Ergebnisse der durchgeführten Studien geben Anhaltspunkte dafür, dass Musikalität durch ein Ein-Faktor-Modell entsprechend Spearmans (1904b) Generalfaktormodell der Intelligenz konzeptualisiert werden kann. Ein Musikalitätskonzept, das zu Gardners (2006) Modell der multiplen, unverbundenen Intelligenzen analog ist, kann wegen der hohen Korrelationen zwischen den Musiktests ausgeschlossen werden. In die Vorüberlegungen zu dieser Arbeit wurden außerdem verschiedene Studien aus der musikpsychologischen Literatur aufgenommen. Eine davon ist die von Stankov und Horn (1980), die drei Faktoren musikalischer Wahrnehmung identifizieren konnten. Dies steht im Gegensatz zum Ergebnis des einzelnen musikalischen *g*-Faktors der vorliegenden Arbeit. Das musikalische Generalfaktormodell ist vereinbar mit dem Musikalitätskonzept, das mittels Wings (1961) *Standardized Tests of Musical Intelligence* gemessen wird. Es widerspricht Seashores (1919b) Auffassung von multiplen Musikalitätsfaktoren, die unabhängig voneinander durch die *Seashore Measures of Musical Talent* (Seashore, 1919a) erhoben werden.

Eine Stärke dieser Studie und allgemein des *LongGold*-Projekts sind die Anwendung der Item-Response-Theorie und die dadurch möglichen modernen adaptiven Testverfahren. Außerdem kann durch die sehr großen Stichproben aus zwei verschiedenen Ländern eine Generalisierung der Ergebnisse angestrebt werden. Von Okada (2018) stammt eine andere Studie, in der die Struktur musikalischer Fähigkeiten an einer deutlich kleineren Stichprobe bestehend aus 165 Personen zwischen 18 und 29 Jahren ($M = 20$ Jahre) untersucht wurde. Sie untersuchte darin die Leistungen in 15 Musiktests (aufgeteilt in Musikwahrnehmungs- und Produktionstests). Die Höhe der Korrelation zwischen diesen Tests ist annähernd vergleichbar mit den in Tabelle 2 berichteten Korrelationen ($r = ,34$ bis $r = ,42$). Durch eine konfirmatorische Faktorenanalyse konnte Okada vier Musikalitätsfaktoren für (1) Tonhöhe, (2) Timing, (3) Wahrnehmung und (4) Produktion identifizieren. Allerdings korrelierten der Tonhöhen- und Timing- sowie der Wahrnehmungs- und Produktionsfaktor stark miteinander ($r = ,67$ bzw. $r = ,71$; vgl. Ellis, 2010, S. 41), sodass man davon ausgehen kann, dass sie dasselbe dahinterliegende Musikalitätskonstrukt messen und das Modell daher auf zwei Faktoren reduziert werden kann. Auch wenn die Ergebnisse der Studie von Okada (2018) durch die Einschränkung dieser hohen Korrelation nicht vollkommen eindeutig sind, geben sie trotzdem Hinweise darauf, dass eine Übertragbarkeit der Ergebnisse unserer Studie, die für Kinder und Jugendliche im Alter von ungefähr zwölf Jahren gelten, auf eine Stichprobe aus älteren Jugendlichen oder Erwachsenen nicht ohne Weiteres möglich ist. Insbesondere ist nicht geklärt, ob das Generalfaktormodell auch für eine Stichprobe aus erwachsenen Profimusiker*innen seine Gültigkeit behält. In Bezug auf Unterschiede zwischen Kindern und Erwachsenen werfen Gembris et al. (1997, Sp. 869) die Frage auf, „ob die Struktur der Musikalität im Laufe des Lebens gleichbleibt oder ob sie sich verändert“. Diese Frage

kann auf Grundlage der Daten, auf denen diese Arbeit basiert, nicht beantwortet werden, da diese nur eine Momentaufnahme aus einem Schuljahr widerspiegeln. Dank des Langzeitdesigns der *LongGold*-Studie ist es jedoch möglich, eine derartige Untersuchung mit denselben Jugendlichen in einigen Jahren zu wiederholen. Erst dann wird sich herausstellen, ob das Generalfaktormodell auch bei älteren Jugendlichen gilt und wie stabil es über einen längeren Zeitraum bleibt.

Finanzierung

Das *LongGold*-Projekt wird durch das Preisgeld des Anneliese-Meier-Forschungspreises der Humboldt-Stiftung gefördert.

Interessenkonflikte

Die Autor*innen haben erklärt, dass keinerlei konkurrierende Interessen bestehen.

Danksagung

Ein großes Dankeschön gebührt Klaus Frieler und Peter Harrison aus dem *LongGold*-Team für die Aufbereitung und Bereitstellung der Daten, Miriam Eisinger und Nicolas Ruth für die Erhebung der Daten in Baden-Württemberg und Hessen sowie Hsin-Rui Lin für seine statistische Fachunterstützung. Selbstverständlich sei allen Schüler*innen für ihre Teilnahme an der *LongGold*-Studie und vor allem unseren Ansprechpartner*innen an den Schulen für die angenehme Zusammenarbeit gedankt.

Ethikerklärung

Die vorliegende Arbeit wurde in Übereinstimmung mit ethischen Prinzipien und Standards durchgeführt. Die Datenerhebung wurde von den Ethikkommissionen der Leibniz Universität Hannover, der Kultusministerien von Baden-Württemberg und Hessen sowie des Goldsmiths Colleges University of London geprüft und genehmigt.

Datenverfügbarkeit

Die Forschungsdaten dürfen gemäß der Genehmigungen der Ethikkommissionen und Kultusministerien nicht veröffentlicht werden. Bei begründetem Interesse besteht jedoch die Möglichkeit, eine Anfrage an die Projektleitung (Prof. Dr. Daniel Müllensiefen) zu stellen, welche im Einzelfall über einen eventuellen Zugang zu den anonymisierten Daten entscheidet.

Literatur

- Beaujean, A. A. (2014). *Latent variable modeling using R: A step-by-step guide*. Routledge.
- Bortz, J., & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Springer.
- Boyle, J. D., & Radocy, R. E. (1987). *Measurement and evaluation of musical experiences*. Schirmer Books.
- Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press.
<https://doi.org/10.1017/CBO9780511571312>
- Chan, Y. W. F., & Kosinski, M. (2015). *ICAR Project Wiki. International Cognitive Ability Resource (ICAR)*. Aufgerufen am 22.08.2019 von <https://icar-project.com/projects/icar-project/wiki>
- Drake, R. M. (1954). *Manual for Drake Musical Aptitude Tests*. University of Chicago Press.

- Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge University Press.
- Fodor, J. (1983). *The modularity of mind*. MIT Press.
- Frey, A. (2012). Adaptives Testen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Aufl., S. 275–293). Springer. https://doi.org/10.1007/978-3-642-20072-4_2
- Gardner, H. (2006). *Multiple Intelligences*. Basic Books. <https://doi.org/10.1080/1047621950070122>
- Gembris, H. (1997). Historical phases in the definition of musicality. *Psychomusicology: Music, Mind, and Brain*, 16(1–2), 17–25. <https://doi.org/10.1037/h0094070>
- Gembris, H., Steinberg, R., & Kormann, A. (1997). Musikalität. In L. Finscher (Hrsg.), *Die Musik in Geschichte und Gegenwart* (2. Aufl., Sachteil Bd. 6, Sp. 867–921). Bärenreiter.
- Gordon, E. E. (1989). *Manual for the Advanced Measures of Music Audiation (AMMA)*. GIA Publications.
- Gordon, E. E. (2004). *Continuing studies in music aptitudes*. GIA Publications.
- Hargreaves, D., & Lamont, A. (2017). *The psychology of musical development*. Cambridge University Press. <https://doi.org/10.1017/9781107281868>
- Harrison, P. M., & Müllensiefen, D. (2018). Development and validation of the Computerized Adaptive Beat Alignment Test (CA-BAT). *Scientific Reports*, 8, Article 12395. <https://doi.org/10.1038/s41598-018-30318-8>
- Harrison, P. M., Collins, T., & Müllensiefen, D. (2017). Applying modern psychometric techniques to melodic discrimination testing: Item response theory, computerized adaptive testing, and automatic item generation. *Scientific Reports*, 7, Article 3618. <https://doi.org/10.1038/s41598-017-03586-z>
- Hothorn, T., Winell, H., Hornik, K., & Zeileis, A. (2021). *Package 'coin'*. Abgerufen am 05.08.2021 von <https://cran.r-project.org/web/packages/coin/coin.pdf>
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- Jensen, A. (1989). *The g factor: The science of mental ability*. Praeger Publishers/Greenwood Publishing Group.
- Kidd, G. R., Watson, C. S., & Gygi, B. (2007). Individual differences in auditory abilities. *The Journal of the Acoustical Society of America*, 122(1), 418–435. <https://doi.org/10.1121/1.2743154>
- Larrouy-Maestri, P., Harrison, P. M. C., & Müllensiefen, D. (2019). The Mistuning Perception Test: A new measurement instrument. *Behavior Research Methods*, 51, 663–675. <https://doi.org/10.3758/s13428-019-01225-1>
- MacGregor, C., & Müllensiefen, D. (2019). The Musical Emotion Discrimination Task: A new measure for assessing the ability to discriminate emotions in music. *Frontiers in Psychology*, 10, Article 1955. <https://doi.org/10.3389/fpsyg.2019.01955>
- MacGregor, C., Ruth, N., & Müllensiefen, D. (2021). *The Musical Emotion Discrimination Task 2: Developing a large set of short musical clips to test the perception of emotions in music* [Vortrag]. 16. International Conference on Music

- Perception and Cognition/11. Triennial Conference of the European Society for the Cognitive Sciences of Music, globale Hubs. Abgerufen am 04.08.2021 von <https://www.youtube.com/watch?v=m9SFFSr-9Us>
- Mackintosh, N. (2011). *IQ and human intelligence* (2. Aufl.). Oxford University Press.
- McGrew, K. S. (2005). *The Cattell-Horn-Carroll Theory of Cognitive Abilities: Past, present, and future*. In D. P. Flanagan & P. L. Harrison (Hrsg.), *Contemporary intellectual assessment: Theories, tests, and issues* (S. 136–181). The Guilford Press.
- Michaelis, K. F. (1805). Über die Prüfung musikalischer Fähigkeiten. *Berlinische Musikalische Zeitung*, 1(56), 222–230.
- Moosbrugger, H., & Kelava, A. (2012). Qualitätsanforderungen an einen psychologischen Test (Testgütekriterien). In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Aufl., S. 7–26). Springer. https://doi.org/10.1007/978-3-642-20072-4_2
- Müllensiefen, D. (2017a, 15.–17. September). *LongGold: Eine Langzeitstudie zur Entwicklung musikalischer Fähigkeiten im Jugendalter* [Vortrag]. 33. Jahrestagung der Deutschen Gesellschaft für Musikpsychologie, Hamburg.
- Müllensiefen, D. (2017b). Messung musikalischer Begabung. In M. Hasselhorn & U. Trautwein (Hrsg.), *Tests und Trends – Jahrbuch der pädagogisch-psychologischen Diagnostik: Bd. 15. Begabungen und Talente* (S. 125–144). Hogrefe.
- Müllensiefen, D., & Elvers, P. (2018, 23.–28. Juli). *Development of musical abilities: Literature review and analysis of cohort study data* [aufgezeichneter Vortrag]. 15. International Conference on Music Perception and Cognition/10. Triennial Conference of the European Society for the Cognitive Sciences of Music, Graz. Abgerufen am 31.07.2019 von <https://www.youtube.com/watch?v=y84q7VnsNOU&feature=youtu.be&t=826>
- Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of nonmusicians: An index for assessing musical sophistication in the general population. *PLOS ONE*, 9(2), Article e89642. <https://doi.org/10.1371/journal.pone.0089642>
- Müllensiefen, D., Harrison, P., Caprini, F., & Fancourt, A. (2015). Investigating the importance of self-theories of intelligence and musicality for students' academic and musical achievement. *Frontiers in Psychology*, 6, Article 1702. <https://doi.org/10.3389/fpsyg.2015.01702>
- Müllensiefen, D., & Hemming, J. (2018). Musikalische Fertigkeiten und ihre Messbarkeit. In A. C. Lehmann & R. Kopiez (Hrsg.), *Handbuch Musikpsychologie* (S. 93–119). Hogrefe.
- Myers, D. G. (2014). Intelligenz. In D. G. Myers (Hrsg.), *Psychologie* (3. Aufl., S. 399–436). Springer. https://doi.org/10.1007/978-3-642-40782-6_11
- Neyer, F. J., & Asendorpf, J. B. (2018). Psychologie der Persönlichkeit. In F. J. Neyer & J. B. Asendorpf (Hrsg.), *Psychologie der Persönlichkeit* (6. Aufl., S. 135–235). Springer. https://doi.org/10.1007/978-3-662-54942-1_4
- Okada, B. M. (2018). *Towards a comprehensive model of musical ability* [Dissertation, University of Maryland]. Digital Repository at the University of Maryland. <https://drum.lib.umd.edu/handle/1903/21621>
- Patel, A. D., & Iversen, J. R. (2014). The evolutionary neuroscience of musical beat perception: The Action Simulation for Auditory Prediction (ASAP) hypothesis. *Frontiers in Systems Neuroscience*, 8, Article 57. <https://doi.org/10.3389/fnsys.2014.00057>

- Peretz, I., & Coltheart, M. (2003). Modularity of music processing. *Nature Neuroscience*, 6, 688–691.
<https://doi.org/10.1038/nn1083>
- Raven, J. (2008). The Raven Progressive Matrices Tests: Their theoretical basis and measurement model. In J. Raven & C. J. Raven (Hrsg.), *Uses and abuses of intelligence: Studies advancing Spearman and Raven's quest for non-arbitrary metrics* (S. 17–68). Royal Fireworks Press.
- Rose, D. (2005). *The ESeC class schema summarized*. Abgerufen am 03.08.2021 von
<https://www.iser.essex.ac.uk/archives/esec/validation-process>
- Rose, D., & Pevalin, D. J. (2001). *The national statistics socio-economic classification. Unifying official and sociological approaches to the conceptualisation and measurement of social class* (ISER Working Papers 2001-04). University of Essex.
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36.
<https://doi.org/10.18637/jss.v048.i02>
- Schaal, N. K., Bauer, A.-K. R., & Müllensiefen, D. (2014). Der Gold-MSI: Replikation und Validierung eines Fragebogeninstrumentes zur Messung Musikalischer Erfahrungheit anhand einer deutschen Stichprobe. *Musicae Scientiae*, 18(4), 423–447. <https://doi.org/10.1177/1029864914541851>
- Seashore, C. E. (1919a). *Manual of instructions and interpretations for Measure of Musical Talent*. Columbia Graphophone Company.
- Seashore, C. E. (1919b). *Beverly educational series: The psychology of musical talent*. Silver, Burdett & Company.
<https://doi.org/10.1037/13031-000>
- Seashore, C. E., Lewis, D., & Saetveit, J. C. (1960). *Manual of instruction and interpretations for the Seashore Measures of Musical Talents* (2nd revision). The Psychological Corporation.
- Shevlin, M., & Miles, J. N. V. (1998). Effects of sample size, model specification and factor loadings on the GFI in confirmatory factor analysis. *Personality and Individual Differences*, 25(1), 85–90.
[https://doi.org/10.1016/S0191-8869\(98\)00055-5](https://doi.org/10.1016/S0191-8869(98)00055-5)
- Spearman, C. (1904a). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1), 72–101. <https://doi.org/10.2307/1412159>
- Spearman, C. (1904b). “General intelligence,” Objectively determined and measured. *The American Journal of Psychology*, 15(2), 201–292. <https://doi.org/10.2307/1412107>
- Stankov, L., & Horn, J. L. (1980). Human abilities revealed through auditory tests. *Journal of Educational Psychology*, 72(1), 21–44. <https://doi.org/10.1037/0022-0663.72.1.21>
- Watson, J. C. (2017). Establishing evidence for internal structure using exploratory factor analysis. *Measurement & Evaluation in Counseling & Development*, 50(4), 232–238. <https://doi.org/10.1080/07481756.2017.1336931>
- Werner, C. S. (2015). *Strukturgleichungsmodelle mit R und lavaan analysieren: Kurzeinführung*. Abgerufen am 31.07.2019 von https://www.psychologie.uzh.ch/dam/jcr:fffff-b371-2797-ffff-ffffeb61aa16/einfuehrung_lavaan_cswerner.pdf

- Werner, C. S., Schermelleh-Engel, K., Gerhard, C., & Gäde, J. C. (2016). Strukturgleichungsmodelle. In N. Döring & J. Bortz (Hrsg.), *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Aufl., S. 945–973). Springer. https://doi.org/10.1007/978-3-642-41089-5_17
- Wing, H. D. (1941). A factorial study of musical tests. *The British Journal of Psychology. General Section*, 31(4), 341–355. <https://doi.org/10.1111/j.2044-8295.1941.tb00999.x>
- Wing, H. D. (1961). *Standardized tests of musical intelligence*. NFER Pub.
- Wing, H. D. (1962). A revision of the Wing Musical Aptitude Test. *Journal of Research in Musical Education*, 10, 743–791.

Appendix

Tabellen

Tabelle A1

Die verwendeten adaptiven Musikalitätstests im Überblick

Test-kürzel	Name (dt.)	Name (engl.)	Item-anzahl	Retest-Reliabilität	empirische Reliabilität	n_{GB}	n_{DE}	Referenz
BAT	Beat-Wahrnehmungstest	Computerised Adaptive Beat Alignment Test	22	,610	,601	622	1164	Harrison und Müllensiefen (2018)
EDT	Emotionsunterscheidungstest	Musical Emotion Discrimination Task	18	,690	—	505	194	MacGregor und Müllensiefen (2019)
MDT	Melodieunterscheidungstest	Melodic Discrimination Test	18	,754	,571	596	1154	Harrison et al. (2017)
MPT	Intonationswahrnehmungstest	Mistuning Perception Test	22	,595	,641	622	1164	Larrouy-Maestri et al. (2019)

Anmerkung. Die Retest-Reliabilität ist die (simulierte) Test-Retest-Korrelation in der jeweiligen Kalibrierungsstudie. Sie unterscheidet sich von der in der Referenz genannten empirischen Reliabilität durch die unterschiedliche Anzahl an für die Berechnung verwendeten Items. Die empirische Reliabilität ist aus den *Standard Errors of the Estimate* (Messfehlern) der Gesamtstichprobe mit dem R-Paket *mirt* (Funktion *empirical_rxx*) berechnet.

Tabelle A2

Normen des Beat-Wahrnehmungstests (BAT), des Melodieunterscheidungstests (MDT) und des Intonationswahrnehmungstests (MPT) nach Alter, Geschlecht und Musikalische Ausbildung (Faktor F3 des Gold-MSI)

Level	Quantil Faktor F3 Musikal. Ausbildung	n_{Level}	BAT		MDT		MPT	
			M	SD	M	SD	M	SD
Subgruppe A: 10 Jahre und weiblich								
1	2,71	65	-1,69	1,34	-1,50	0,98	-0,93	1,21
2	3,86	53	-1,31	1,23	-1,15	0,95	-0,31	1,19
3	6,14	52	-1,16	1,30	-1,16	1,05	-0,31	1,16
gesamt	—	170	-1,41	1,31	-1,29	1,00	-0,55	1,22

Level	Quantil Faktor F_3 <i>Musikal. Ausbildung</i>	n_{Level}	<i>BAT</i>		<i>MDT</i>		<i>MPT</i>	
			<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Subgruppe B: 10 Jahre und männlich								
1	2,29	61	-1,82	1,08	-1,65	0,98	-1,14	1,21
2	3,14	52	-1,98	1,14	-1,62	1,03	-0,95	1,20
3	6,14	56	-1,27	1,24	-1,24	1,20	-0,78	1,13
gesamt	—	169	-1,69	1,19	-1,50	1,08	-0,96	1,18
Subgruppe C: 11 Jahre und weiblich								
1	2,29	64	-1,79	1,46	-1,74	0,93	-0,59	1,05
2	3,14	51	-1,53	1,18	-1,27	1,01	-0,53	1,36
3	3,71	55	-1,49	1,15	-1,41	1,10	-0,29	0,96
4	4,43	51	-1,36	1,43	-0,96	1,24	-0,53	1,18
5	6,29	54	-1,16	1,33	-0,83	1,19	0,02	0,96
gesamt	—	275	-1,48	1,33	-1,26	1,14	-0,39	1,12
Subgruppe D: 11 Jahre und männlich								
1	2,00	64	-1,95	1,28	-1,66	1,01	-1,36	1,26
2	2,86	68	-1,64	1,36	-1,76	1,02	-1,00	1,30
3	4,00	58	-1,19	1,36	-1,34	1,08	-0,64	1,05
4	5,86	56	-1,18	1,35	-1,15	1,15	-0,40	1,11
gesamt	—	246	-1,51	1,37	-1,50	1,08	-0,87	1,24
Subgruppe E: 12 Jahre und weiblich								
1	2,43	61	-1,18	1,31	-1,25	0,90	-0,36	1,07
2	3,43	57	-0,93	1,20	-1,14	1,18	-0,10	1,01
3	4,15	49	-0,87	1,06	-0,60	1,15	0,21	0,78
4	6,29	55	-0,56	1,30	-0,55	1,14	0,12	1,13
gesamt	—	222	-0,89	1,24	-0,90	1,13	-0,05	1,03
Subgruppe F: 12 Jahre und männlich								
1	2,14	54	-1,61	1,24	-1,77	1,16	-0,82	1,10
2	3,29	49	-1,76	1,31	-1,58	1,25	-0,80	1,14
3	6,43	50	-1,16	1,46	-1,21	1,28	-0,23	1,10
gesamt	—	153	-1,51	1,35	-1,53	1,24	-0,62	1,14
Subgruppe G: 13 Jahre und weiblich								
—	—	92	-0,65	1,17	-0,84	1,10	0,21	1,01
Subgruppe H: 13 Jahre und männlich								
—	—	54	-0,91	1,20	-1,14	1,12	-0,67	1,17
Subgruppe I: 14 Jahre und weiblich								
—	—	62	-0,48	1,09	-0,80	1,24	0,06	1,12

Anmerkung. Die Werte des *BATs*, *MDTs* und *MPTs* sind teilweise negativ, weil die Metrik gemäß der Item-Response-Theorie nur Werte von -4 bis $+4$ umfasst und 0 dem Mittelwert der erwachsenen Kalibrierungsstichprobe entspricht. Die Anzahl k der Substichproben ist abhängig von $n_{\text{Subgruppe}}$. Die oberen Grenzwerte der Variable *Musikalische Ausbildung*, nach der die Einteilung vorgenommen wurde, sind in Spalte 2 abgetragen. Sie entsprechen den $\frac{1}{k}$, $\frac{2}{k}$ usw. Quantilen von F_3 . In Spalte 3 sind die Größen der letztendlichen Substichproben aufgeführt. Wenn die Subgruppe nicht größer als $n_{\text{Subgruppe}} = 100$ war (Subgruppe G, H und I), gibt es nur ein Level der Variable *Musikalische Ausbildung*. Das Minimum von Level 1 hat für alle Subgruppen den Wert 1.

Explorative Faktorenanalyse

Das Kaiser-Meyer-Olkin-Kriterium (KMO), die Measures of Sampling Adequacy (MSA) und das Ergebnis aus Bartlett's Test auf Sphärität lassen darauf schließen, dass die Daten für die vier Musikalitätstests der britischen Stichprobe für eine Faktorenanalyse geeignet sind. Zunächst wird der Scree-Plot betrachtet (siehe [Abb. A1](#)). Gemäß dem Kaiser-Guttman-Kriterium sind nur Faktoren mit einem Eigenwerte über 1 als bedeutsam einzustufen (Bortz & Schuster, 2010, S. 415; in

Abb. A1 durch die horizontale Linie markiert). Der Eigenwert des ersten Faktors beträgt 1,41 und ist daher größer als eins (vgl. Abb. A1). Alle anderen Faktoren weisen einen Eigenwert kleiner als eins auf. Daher wird in der Faktorenanalyse die Anzahl der zu extrahierenden Faktoren auf eins gesetzt. Eine explorative Faktorenanalyse mit obliquen Oblimin-Rotation und Maximum-Likelihood-Methode extrahiert einen Faktor mit 35,2% Varianzaufklärung. Die Variablen *BAT*, *EDT*, *MDI* und *MPT* haben Faktorladungen von 0,560, 0,551, 0,652 und 0,606. Diese sind annähernd so groß wie die Ladungen aus Modell 1.

Abbildung A1

Scree-Plot

