

Mary Laqua<sup>a</sup>

## Der Mozart-Effekt – Wahrheit oder Mythos?<sup>1</sup>

Als in einer praxisorientierten wissenschaftlichen Disziplin Arbeitende, stehen Musiktherapeut:innen vor der Herausforderung, wichtige behandlungsbezogene Entscheidungen zu treffen. Neben der Berücksichtigung klinischer Erfahrungen und der Werte und Bedürfnisse von Klient:innen, besteht eine evidenzbasierte Vorgehensweise darin, dass Musiktherapeut:innen bei behandlungsbezogenen Entscheidungen wissenschaftliche Erkenntnisse einbeziehen, diese kritisch bewerten und sich davon leiten lassen. Die Veröffentlichung einer Studie in einer wissenschaftlichen, von Fachkolleg:innen begutachteten Zeitschrift bedeutet jedoch nicht, dass die Studie gut durchgeführt wurde, die Daten korrekt analysiert und Ergebnisse richtig interpretiert wurden (Steinberg & Luce, 2005). Bevor man einen Artikel als handlungsleitend bezeichnet, sollte man ihn kritisch prüfen, um die methodische Qualität und die Aussagekraft der Schlussfolgerungen zu beurteilen. Bei der kritischen Bewertung einer Studie, die darauf abzielt, die Wirksamkeit einer therapeutischen Intervention zu testen, sind vier Kriterien von zentraler Bedeutung: (1) interne Gültigkeit, (2) Messvalidität und Verzerrung, (3) Gültigkeit der statistischen Schlussfolgerung und (4) externe Validität (Rubin & Bellamy, 2022).

Um den Prozess der kritischen Bewertung der empirischen Aussagekraft zu veranschaulichen, wird auf eine bekannte musikbezogene und angeblich wissenschaftliche Theorie eingegangen, den sogenannten *Mozart-Effekt*.

### Der Mozart-Effekt

Der Mozart-Effekt tauchte erstmals in einem einseitigen Artikel in der renommierten Fachzeitschrift *Nature* unter dem Titel »Music and spatial task performance« auf (Rauscher, Shaw & Ky, 1993). Das Studienergebnis war, dass 36 Studierende nach dem Anhören von 10 Minuten der Mozart-Sonate für zwei Klaviere in D-Dur, KV 448, ihre räumliche Intelligenz vorübergehend für 12 bis 15 Minuten steigern konnten. Das methodische Vorgehen wurde von den Autor:innen folgendermaßen beschrieben:

---

a Redakteurin FZS Musiktherapeutische Umschau, Holzkirchen

Im Experiment erhielten 36 Studierende jeweils drei Reihen von Standardaufgaben zum räumlichen Denken. Jeder Aufgabe gingen die folgenden Hörbedingungen voraus:

1. Eine Musikgruppe, in der jede Testperson 10 Minuten lang Mozarts Sonate KV 448 anhörte.
2. Eine Entspannungsgruppe, in der jede Testperson 10 Minuten lang Entspannungsanweisungen vom Tonträger anhörte.
3. Eine Stillegruppe, in der jede Testperson 10 Minuten lang schweigend dasaß.

Die 36 Studierenden nahmen an allen drei Hörbedingungen teil. Unmittelbar nach jeder Hörbedingung wurde einer von drei Tests zum abstrakten Denken aus der Stanford-Binet-Intelligenzskala (Thorndike, Hagen & Sattler, 1986) durchgeführt. Die Aufgaben zum abstrakten räumlichen Denken bestanden aus einem Musteranalysetest, einem Multiple-Choice-Matrizen-Test und einem Multiple-Choice-Test zum Papierfalten und -schneiden. Den Autor:innen zufolge korrelierten diese drei Tests auf einem Signifikanzniveau von  $p = 0,01$ , wodurch sie als gleichwertige Maße für abstrakt-räumliches Denken eingestuft wurden.

Die Rohwerte wurden berechnet, indem die Anzahl der nicht bestandenen Aufgaben von der höchsten Anzahl der durchgeführten Aufgaben abgezogen wurde. Die Autor:innen verwendeten dann den Stanford-Binet Standard Age Score mit einem Mittelwert von 50 und einer Standardabweichung von 8, um die Rohwerte in Standard-Alters-Werte umzurechnen. Die IQ-Äquivalente wurden zunächst berechnet, indem jeder Standard-Alters-Wert mit 3 (die Anzahl der von der Stanford-Binet-Skala für die Berechnung des IQ erforderlichen Untertests) multipliziert wurde. Rauscher et al. (1993) verwendeten dann die Umrechnungstabelle der Skala, die auf einen Mittelwert von 100 und eine Standardabweichung von 16 ausgelegt ist, um Standard-Alters-Werte-IQ-Äquivalente zu erzielen. Die Standard-Alters-Werte für die drei Gruppen waren wie folgt:

1. Die Musikgruppe erzielte einen mittleren Standard-Alters-Wert von 57,56.
2. Die Entspannungsgruppe erzielte einen mittleren Standard-Alters-Wert von 54,61.
3. Die Stillegruppe erzielte einen mittleren Standard-Alters-Wert von 54,00.

Um die Ergebnisse besser interpretieren zu können, »übersetzten« die Autor:innen sie in räumliche IQ-Werte von 119, 111 bzw. 110. Somit lag der IQ der Teilnehmer:innen, die an der Musikbedingung teilnahmen, um 8 bis 9 Punkte über den IQ-Werten in den anderen zwei Bedingungen.

Die kritische Bewertung der Methodik und der Ergebnisse hat jedoch Zweifel an der empirischen Validität der zugrundeliegenden wissenschaftlichen Befunde zum Mozart-Effekt aufkommen lassen (Steele, Bass, & Crook, 1999; Chabris, 1999; Steele, 2000; Fudin & Lembessis, 2004). Dazu gehören Fragen hinsichtlich des Studiendesigns, der Bewertung der Stanford-Binet-Intelligenzskala und der Validität der IQ-Messung. Vor diesem Hintergrund wird die Studie im Folgenden anhand der oben genannten vier Kernpunkte methodisch überprüft.

## 1. Interne Gültigkeit

Eine Studie zur Wirksamkeit einer Intervention hat insofern interne Gültigkeit, als es ihr Studiendesign ermöglicht, zu überprüfen, ob das beobachtete Ergebnis tatsächlich die Wirksamkeit oder auch Unwirksamkeit der Intervention widerspiegelt und nicht etwa eine andere Erklärung dafür vorliegt (Rubin & Bellamy, 2022). Das Forschungsdesign von Rauscher et al. aus dem Jahr 1993

weist aufgrund einer unklaren Beschreibung der Methodik gravierende Mängel auf. Sie berichteten: »Die Studierenden erhielten jeweils 3 Reihen von Standard-IQ-Aufgaben zum räumlichen Denken«, »36 Studierende nahmen an allen 3 Hörsituationen teil« und »einer von 3 Tests zum abstrakten Denken aus der Stanford-Binet-Intelligenzskala wurde nach jeder Hörsituation durchgeführt« (S. 611). Die Einzelheiten eines Forschungsdesigns, das auf dem ersten Zitat basiert, sind unklar, da die Stanford-Binet-Skala nur *eine* Reihe von Tests zum abstrakten/visuellen Denken umfasst. Die Aussage zur Verwendung von 3 *Aufgabenreihen* enthält keine Details zum Verfahren, wie zwei neue Aufgabenreihen in ihrem Schwierigkeitsgrad an die einzige standardisierte Reihe aus der Stanford-Binet-Intelligenzskala angepasst wurden.

Obwohl in der Studie angegeben wurde, dass Standard-Alters-Werte aus der Stanford-Binet-Intelligenzskala verwendet wurden, um das räumliche Denkvermögen der Studierenden nach dem Anhören von Mozart zu bewerten, handelt es sich tatsächlich um den *mittleren Standard-Alters-Wert von 3 verschiedenen Gruppen mit jeweils 12 Teilnehmenden*. Die gleiche Beobachtung gilt für den mittleren Standard-Alters-Wert für die Bedingungen »Entspannung« und »Stille«. Daraus folgt, dass die angegebenen räumlichen IQ-Werte nicht auf den Ergebnissen von 3 Subtests pro Teilnehmer:in und Hörbedingung basierten, sondern nur auf *einem Test* pro Teilnehmer:in und Hörbedingung (Fudin & Lembessis, 2004).

## 2. Messvalidität und Verzerrung

Hiermit stellt sich die zentrale Frage: Wurde das Ergebnis auf stichhaltige und unverzerrte Weise gemessen? Rauscher et al. (1993) verwendeten 3 Subtests aus der Stanford-Binet-Skala, die den Bereich »abstraktes/visuelles Denken« umfassen, nämlich Musteranalyse, Matrizen sowie Papierfalten und -schneiden. In der Anwendung sind individuelle Subtest-Standard-Alters-Werte erforderlich, um einen Bereichsstandard-Alters-Wert zu erhalten. Bei der Studie mussten daher die Rohwerte für Musteranalyse, Matrizen und Papierfalten und -schneiden zunächst in Subtest-Standard-Alters-Werte umgewandelt werden. Diese Werte wurden dann addiert und in einen Standard-Alters-Wert für den Bereich abstraktes/visuelles Denken umgewandelt. Rauscher et al. haben *keine* summierten Standard-Werte für die einzelnen Subtests für jede Hörbedingung angegeben und konnten daher die mittleren Standard-Alters-Werte für die einzelnen Subtests nicht direkt in einen Bereichs-Standard-Alters-Wert (ihren »räumlichen IQ-Wert«) umrechnen. Zusammengefasst: Die einzigen Werte, die von Rauscher et al. (1993) hinsichtlich der Validität der Leistungsmessungen präsentiert und statistisch analysiert wurden, waren die mittleren Standard-Alters-Werte der verschiedenen Hörgruppen (Fudin & Lembessis (2004). Zudem bezeichneten sie die Standard-Alters-Werte für abstrakte/visuelle Denkaufgaben irreführenderweise als »Standard-Alters-Werte-IQ-Äquivalente«, »räumliche IQ-Werte« oder einfach nur »IQs« (1993, S. 611).

Damit die Teilnehmenden bei einer Aufgabenstellung tatsächliche Verbesserungen zeigen können, muss ein Basiswert (*baseline measurement*) zum Vergleich der 3 Gruppen vorliegen. Eine Baseline ist ein fest definierter Ausgangspunkt oder eine Vergleichsbasis, die als Referenz dient, um Fortschritte und Veränderungen zu messen. In dieser Studie hätte dies dadurch erreicht werden können, dass alle 36 Teilnehmenden bereits vor den Versuchsbedingungen identische Aufgaben aus dem Stanford-Binet-Intelligenztest zum Papierfalten und -schneiden abgelegt hätten. Diese Werte werden verwendet, um anhand der Mittelwerte die 3 Gruppen miteinander zu ver-

gleichen. Nach dem Experiment werden dann die Baseline-Werte mit späteren Messungen verglichen. Rauscher et al. (1993) haben jedoch keine Basiswerte gemessen, die belegen, dass das Hören von Mozarts Musik die Leistung bei einer einzelnen Aufgabe verbessert oder steigert (Fudin & Lembessis, 2004).

Ein weiteres Risiko liegt in der Veröffentlichungsverzerrung (*publication bias*), was bedeutet, dass Studien mit positiven Ergebnissen eine größere Wahrscheinlichkeit haben, von Fachzeitschriften angenommen zu werden. Dies ist ein allgemeines Phänomen und führt zu einer Überschätzung eines tatsächlichen Effekts. Ausgehend von den Erkenntnissen von Rauscher et al. (1993, 1994) wurde der Mozart-Effekt fast ausschließlich im Zusammenhang mit IQ-Werten betrachtet. Dabei wurde jedoch nur *eine* Art von Intelligenz getestet, nämlich das räumliche Vorstellungsvermögen. Die hiermit verbundene Veröffentlichungsverzerrung hat scheinbar zu einem ungenauen und überhöhten Verständnis des Mozart-Effekts beigetragen und einen Mythos in der öffentlichen Diskussion aufrechterhalten, was zu Behauptungen über langanhaltende kognitive Vorteile geführt hat.

### 3. Gültigkeit statistischer Schlussfolgerungen

Das Thema statistischer Signifikanz ist komplex und oft schwer zu begreifen (Bergmann, 2018). Eine Frage, die man sich zu Ergebnissen bezogen auf Wirksamkeit stellen sollte, lautet: »Wie hoch ist die Wahrscheinlichkeit, dass die offensichtliche Wirksamkeit oder fehlende Wirksamkeit auf statistischen Zufall zurückzuführen ist?« (Rubin & Bellamy, 2022). Diese Frage wird gestellt, weil eine mangelhafte Randomisierung zu Gruppenunterschieden führen und der Grund dafür sein kann, dass bei einer Gruppe bessere Ergebnisse erzielt werden als bei einer anderen.

Um statistische Schlussfolgerungen zu ziehen, sind angemessene Stichprobengrößen, geeignete statistische Methoden und die Kontrolle von verzerrenden Faktoren erforderlich, um Zusammenhänge genau zu erkennen und Fehler wie die Verwechslung von Zufall und tatsächlicher Wirkung zu vermeiden. Die ursprüngliche Studie von Rauscher et al. (1993) hat eine geringe statistische Aussagekraft, da sie mit lediglich 36 Studienteilnehmer:innen nicht über eine ausreichende Stichprobengröße verfügten, um eine tatsächliche Wirkung zuverlässig nachzuweisen. Solche *underpowered studies* sind problematisch, da sie zu verzerrten Schlussfolgerungen führen (Crutzen & Peters, 2017).

### 4. Externe Validität

Bei der externen Validität wird gefragt, ob die Ergebnisse auf andere Gruppen und Kontexte oder auf längere Zeiträume übertragen werden können. Ein vorübergehender oder geringer Effekt kann die externe Validität einer Studie einschränken, da die Ergebnisse möglicherweise nicht auf andere Kontexte oder längere Zeiträume übertragbar sind. Die in der Studie von Rauscher et al. (1993) berichteten Verbesserungen waren von kurzer Dauer (etwa 12 bis 15 Minuten), was ihre externe Validität einschränkt. Die alleinige Auswahl von Studierenden als Testpersonen wirkt sich ebenfalls negativ auf die externe Validität aus, da sie es schwierig macht, die Studienergebnisse auf breitere Bevölkerungs- und Altersgruppen sowie Umgebungen und Situationen zu übertragen.

Um eine Generalisierung von Effekten über die Zeit nachzuweisen, ist eine weitere Messung mit Abstand zur Intervention (follow-up measurement) notwendig. In einer nachfolgenden Studie von 1994 strebten Rauscher et al. an, die Ergebnisse ihrer Studie von 1993 zu wiederholen so-

wie die Wirkung von mehrmaligem Anhören von Mozarts Sonate für zwei Klaviere KV 448 auf Tests zum räumlichen Denken zu erforschen. Studierende von derselben Universität wurden noch einmal getestet – dieses Mal über einen Zeitraum von 5 Tagen. Von 84 Teilnehmenden haben 79 die 5 Forschungstage abgeschlossen. Zusätzlich zu einer Mozartgruppe und einer Stillegruppe kam eine gemischte Gruppe dazu, die der Erforschung der Wirkungen von sich wiederholender Musik auf räumliches Denken (Papierfalten und -scheiden) diente. Die Wirkung von Mozarts Musik auf das Kurzzeitgedächtnis wurde auch erfasst. Hiermit wird es deutlich, dass die Forschungsdesigns der beiden Studien unterschiedlich waren, was einen Vergleich der Ergebnisse erschwert. (Fudin & Lembessis, 2004).

Je mehr Replikationen einer Studie mit ähnlichen Ergebnissen vorliegen, desto höher ist die Wahrscheinlichkeit der externen Validität und desto wahrscheinlicher ist es, dass die Ergebnisse auf einen breiteren Kontext verallgemeinert werden können. Spätere Untersuchungen deuteten darauf hin, dass alle beobachteten Effekte auf das räumliche Denken wahrscheinlich auf eine vorübergehende Verbesserung der Stimmung und Erregung sowie Unterschiede in den Musikvorlieben zurückzuführen sind und nicht speziell auf Mozarts Sonate für zwei Klaviere KV 448 (Jones, West & Estell, 2006; Steele, 2000; Steele, Bass & Crook, 1999; Chabris, 1999; Nantais & Schellenberg (1999).

Angesichts dieser Tatsachen kommen Oberleiter und Pietschnig (2023) in ihrer in *Nature Scientific Reports* veröffentlichten Metaanalyse zu dem Schluss, dass unbegründete Autorität, fehlerhafte Studien und mangelnde Transparenz in der Berichterstattung den Mythos des Mozart-Effekts offenbar aufrechterhalten.

## Aktuelle Relevanz

Trotz seiner nachweislichen wissenschaftlichen Untauglichkeit lebt der Mythos vom Mozart-Effekt weiter, da die anhaltende Faszination für Mozart als Katalysator für Intelligenz breites öffentliches Interesse gefunden hat. Welche Eltern würden nicht gerne die kognitiven Fähigkeiten ihres Kindes fördern? Als Reaktion auf diesen Wunsch hat eine von den Medien angeheizte weit verbreitete Fehlinterpretation eine Milliarden-Dollar-Industrie von Büchern und Tonträgern hervorgebracht, die verspricht, kognitive Fähigkeiten zu verbessern. Dies spiegelt das große allgemeine Interesse an der Wirkung von Musik auf das Gehirn wider, die nach wie vor systematisch erforscht wird.

Musiktherapeut:innen sehen sich mit Fakten, Theorien und Mythen über den Einfluss von Musik auf viele Aspekte des menschlichen Lebens konfrontiert, sei es auf die körperlichen, emotionalen, kognitiven oder sozialen Bedürfnisse des Einzelnen. Neben der Berücksichtigung klinischer Erfahrungen und der Werte und Bedürfnisse von Klient:innen besteht eine evidenzbasierte Vorgehensweise darin, dass Musiktherapeut:innen bei behandlungsbezogenen Entscheidungen wissenschaftliche Erkenntnisse einbeziehen, diese kritisch bewerten und sich davon leiten lassen. Der Schlüssel zu diesem Prozess liegt darin, Fragen zu stellen und Antworten zu finden, um interne Gültigkeit, Messvalidität und Verzerrung, statistische Schlussfolgerungsgültigkeit und externe Validität zu erfassen.

## Literatur

- Bergmann, T. (2018). Vorsicht, signifikant! *Musiktherapeutische Umschau* 39(2), 188–191.
- Chabris, C. F. (1999). Prelude or requiem for the »Mozart effect? *Nature*, 400(6747), 826–827.
- Crutzen, R. & Peters, G. Y. (2017). Targeting Next Generations to Change the Common Practice of Underpowered Research. *Front. Psychol.*, 13(8),1184. <https://doi.org/10.3389/fpsyg.2017.01184>
- Fudin, R. & Lembessis, E. (2004). The Mozart effect: questions about the seminal findings of Rauscher, Shaw, and colleagues. *Percept Mot Skills*, 98(2), 389–405. <https://doi.org/10.2466/pms.98.2.389-405>
- Jones, M. H., West, S. D., & Estell, D. B. (2006). The Mozart effect: Arousal, preference, and spatial performance. *Psychology of Aesthetics, Creativity, and the Arts*, 5(1), 26–32. <https://doi.org/10.1037/1931-3896.S.1.26>
- Nantais, K. M., & Schellenberg, E. G. (1999). The Mozart effect: an artifact of preference. *Psychological science*, 10, 370–373.
- Oberleiter, S. & Pietschnig, J. (2023). Unfounded authority, underpowered studies, and non-transparent reporting perpetuate the Mozart effect myth: A multiverse meta-analysis. *Scientific Reports*, 13, 3175. <https://doi.org/10.1038/s41598-023-30206-w>
- Rauscher, F. H., Shaw, G. L. & Ky, K. N. (1993). Music and spatial task performance. *Nature* 365(6447). <https://doi.org/10.1038/365611a0>
- Rauscher, F. H., Shaw, G. L., Levine, L. J., Ky, K. N. & Wright, L. (1994). *Music and spatial task performance: a causal relationship*. Presented at the meeting of the American Psychological Association, Los Angeles, CA.
- Rubin, A. & Bellamy, J. (2022). *Practitioner's Guide to Using Research for Evidence-Informed Practice* (3rd ed.). Hoboken, NJ: John Wiley & Sons.
- Steele, K. M. (2000). Arousal and mood factors in the »Mozart effect.« *Perceptual and Motor Skills*, 91, 188–190.
- Steele, K. M., Bass, K. E. & Crook, M.D. (1999). The mystery of the Mozart effect: failure to replicate: *Psychological Science*, 10, 366–369.
- Steinberg, E. P. & Luce, B. R. (2005). Evidenced based? Caveat emptor! *Health Affairs*, 24(1), 80–92. <https://doi.org/10.1377/hlthaff.24.1.80>
- Thorndike, R. L., Hagen, E. P. & Sattler, J. M. (1986). *The Stanford-Binet Intelligence Scale: guide for administering and scoring* (4th ed.). Chicago, IL: Riverside.

1 Der vorliegender Beitrag wurde von der Autorin in englischer Sprache geschrieben, mithilfe von DeepL.com ins Deutsche übertragen und durch die Autorin und Thomas Bergmann nachbereitet.



**Mary Laqua, Holzkirchen (Obb.)**  
mary.laqua@musiktherapie.de

Dipl. Musiktherapeutin, DMtG, Neurologic Music Therapy®, langjährige musiktherap. Erfahrung in der stat. Altenpflege sowie in der neurolog. Frührehabilitation, Redakteurin der FZS Musiktherapeutische Umschau.