The Associations Between Music Training, Musical Working Memory, and Visuospatial Working Memory: An Opportunity for Causal Modeling

Sebastian Silas
Hochschule für Musik, Theater und Medien, Hanover, Germany

Daniel Müllensiefen
Goldsmiths University of London, London, United Kingdom

Rebecca Gelding
Macquarie University, Sydney, Australia

Klaus Frieler
Max Planck Institute for Empirical Aesthetics, Frankfurt/M., Germany

Peter M. C. Harrison
University of Cambridge, Cambridge, United Kingdom

Prior research studying the relationship between music training (MT) and more general cognitive faculties, such as visuospatial working memory (VSWM), often fails to include tests of musical memory. This may result in causal pathways between MT and other such variables being misrepresented, potentially explaining certain ambiguous findings in the literature concerning the relationship between MT and executive functions. Here we address this problem using latent variable modeling and causal modeling to study a triplet of variables related to working memory: MT, musical working memory (MWM), and VSWM. The triplet framing allows for the potential application of d-separation (similar to mediation analysis) and V-structure search, which is particularly useful since, in the absence of expensive randomized control trials, it can test causal hypotheses using cross-sectional data. We collected data from 148 participants using a battery of MWM and VSWM tasks as well as a MT questionnaire. Our results suggest: 1) VSWM and MT are unrelated, conditional on MWM; and 2) by implication, there is no far transfer between MT and VSWM without near transfer. However, the data are unable to distinguish an unambiguous causal structure. We conclude by discussing the possibility of extending these models to incorporate more complex or cyclic effects.

Received: March 19, 2021, accepted November 20, 2021.

Key words: working memory, causal modeling, music training, musical working memory, visuospatial working memory

The literature regarding music training (MT) and other nonmusical cognitive faculties, like working memory (WM), is somewhat ambiguous, with competing evidence and explanations for the same or similar sets of variables. For instance, several authors have documented a positive correlation between MT and various WM capacities, such as visuospatial working memory (VSWM; Talamini et al., 2016, 2017). This positive correlation has been explained via two general arguments that are often assumed to be incompatible or competing with one another.

On one side, taking approaches related to genetics, it can be argued that prior dispositions, subserved by genotypic variation, are the primary cause for the positive correlation (i.e., those with better nonmusical WM capacities, in the first place, are predisposed towards developing acute musical memory faculties). This *pre-existing dispositions hypothesis* is rooted in the idea that pre-existing dispositions (which could be due to genetics but also other developmental factors, e.g., socioeconomic background) play the fundamental role in simultaneously guiding the acquisition of musical and nonmusical skills, hence explaining the positive correlations between MT and nonmusical cognitive faculties (Mosing et al., 2015; Mrazik & Dombrowski, 2010; Plomin et al., 2016; Tan et al., 2014; Vinkhuyzen et al., 2009).

On the other hand, particularly driven by the field of neuroscience, it is suggested that MT not only improves musical memory faculties, but also potentially other related WM faculties, such as VSWM, via experience-driven plasticity (Bergman Nutley et al., 2014; George &

Music Perception, volume 39, issue 4, pp. 401–420, issn 0730-7829, electronic issn 1533-8312. DOI: https://doi.org/10.1525/mp.2022.39.4.401

Coch, 2011). This *experience-driven hypothesis* is rooted in the idea that musical abilities are primarily cultivated through training (Ericsson et al., 1993; Ericsson & Moxley, 2013), which, in turn, can have far transfer effects to nonmusical domains (Bigand & Tillmann, 2021; Patel, 2011), perhaps through a profound experience-driven restructuring of the cortex (Bangert & Altenmüller, 2003; Münte et al., 2002; Seither-Preisler et al., 2014).

CAUSALITY AND MUSIC TRAINING

Clearly, the scenarios described above offer alternative causal explanations for the same variables. Since a causal claim bestows a mechanism upon related phenomena, it is a serious affair that should be undertaken with the utmost diligence. Yet, as Schellenberg's (2020) recent literature review demonstrated, when assessing whether MT can have "far transfer" effects to nonmusical domains, researchers have repeatedly committed the error of making (or implying) causal inferences based on correlational data. These biases were found to be systematic: A) positive associations between MT and nonmusical constructs tended to interpret MT as the antecedent, and B) such misinterpretive errors occurred more often in the neuroscience literature, which was more likely to interpret the positive associations as a result of experience-driven plasticity and neglect behavioral genetics approaches. In closing, Schellenberg (2020) notes what is at stake by misrepresenting the causal direction: "it is a disservice [. . .] to offer false hope, wittingly or otherwise, to the public, educators, and other researchers" (p. 479). Hence, given the history of misinterpretation, it is especially important to be cautious with correlational data when studying MT and its often supposed "effects."

LONGITUDINAL DESIGN IN MUSIC TRAINING RESEARCH

In an appropriately designed study, longitudinal research is one way to mitigate the limitations of cross-sectional data, since ascertaining the temporal precedence of certain factors through repeated measurements offers a strong argument for causal inference. The few existing longitudinal studies relevant to our inquiry have found that long-term MT could have positive effects on cognitive and sensorimotor functions (James et al., 2020), implement functional cortical changes (Seither-Preisler et al., 2014), generate small general intellectual benefits (Schellenberg, 2004), and positively influence language development (Lorenzo et al., 2014). Correspondingly, some authors argue their results illustrate that MT has far transfer effects whereby acquired musical skills transfer to general and unrelated domains (e.g., numeracy, academic achievement; Hille & Schupp, 2015; Williams et al., 2015).

However, some of the methodological designs still suffer from a neglect of inherited/pre-existing dispositions approaches, and hence, other longitudinal research has urged more cautious conclusions. For example, MT may not independently contribute to improved academic achievement once certain variables (like IQ and academic performance pre-training) are adjusted for (Yang et al., 2014). Clearly, more longitudinal research is needed to establish a consensus on whether MT can indeed have far transfer effects. However, it will likely take many more years to achieve this based on sufficiently powered longitudinal studies. In the meantime, we attempted to find a way to study MT with cross-sectional data without committing the errors Schellenberg (2020) warned against.

THE MISSING LINK?

We suggest that one reason for the ambiguous findings described above may be that many studies that assess MT and nonmusical variables do not include tests of musical memory in their design (e.g., Bailey & Penhune, 2012; Diaz Abrahan et al., 2019; Schellenberg, 2004). Consequently, when assessing MT alongside a nonmusical variable, one is either assessing a bivariate relationship or not considering musical memory in the variable set. Yet, it seems likely that musical memory faculties are highly relevant variables, and if involved in the measured variable set, could result in different inferences being made. Hence, we decided to design a study that can account for a role of musical memory and, consequently, allow for an intermediary effect to be discovered. Focusing on one such relevant variable that seems to be a likely intermediate between MT and general WM, we characterize musical working memory (MWM) as a domain-specific working memory faculty that supports the temporary retention and manipulation of musical stimuli (e.g., musical notation) in order to perform musical tasks (e.g., sight-reading; Berz, 1995). Those who have undergone intense music training, often referred to as "musicians," should have larger MWM (and potentially other WM) capacities than those with less training (Okada & Slevc, 2018; Talamini et al., 2017).

Alongside MWM and MT, we also focus on VSWM as a representative aspect of nonmusical WM. VSWM, also known as the visuospatial sketchpad, is a component of Baddeley's (2000) multicomponent model of WM, arguably the most widely accepted WM framework (Conway et al., 2013). We focus on VSWM

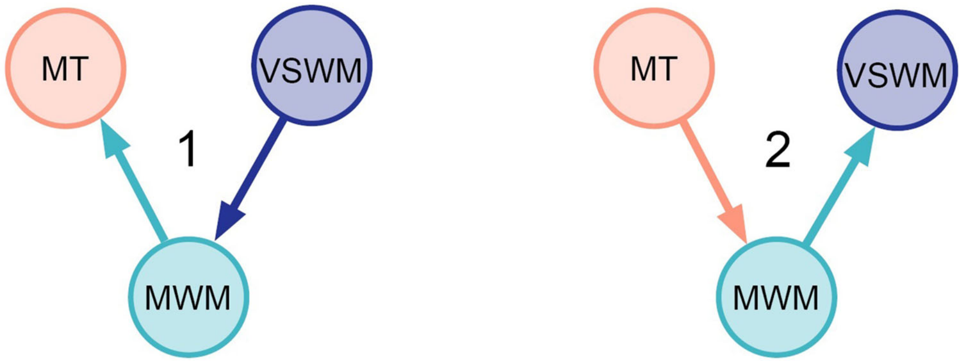

FIGURE 1. Competing hypotheses for the causal structure of music training (MT), musical working memory (MWM), and visuospatial working memory (VSWM).

because it is clearly distinct from the auditory components of MWM, while simultaneously being a strong predictor of performance in important life domains (Anaya et al., 2017; Li & Geary, 2017; Pham & Hasson, 2014). It sometimes also positively correlates with MT (Talamini et al., 2016).

By framing the problem as a triplet of variables, our approach invites the use of informative, and also relatively novel, methodologies.

CORRELATION = CAUSATION? CAUSAL INFERENCE IN STATISTICS

As Schellenberg (2020) noted, researchers need to be very careful about making causal misinterpretations, which seem particularly endemic to studies of MT. Yet, fortunately, it is not strictly true that correlations can *never* imply causation. Recent developments in statistical theory and practice have carefully documented the conditions under which causality may be inferred from cross-sectional data (i.e., via correlations). The *causal inferences in statistics* field has grown substantially over recent years due to the work of Pearl (2000), Spirtes et al. (2000), Morgan (2013), Imbens and Rubin (2015) and many others. Pearl has repeatedly argued that "the mantra correlation does not imply causation should give way to some correlations do imply causation" (Pearl & Mackenzie, 2018, p. 75), because causal models generate testable implications that can be evaluated statistically. This provides a direct answer to the question in Schellenberg's (2020) title, "*Correlation = Causation?*" Yes, if (and only if) models are based on defendable causal assumptions, and the axioms of causal analysis are stringently followed (Pearl, 2000), it is possible to investigate causal hypotheses with cross-sectional data. Hence, using data-driven modeling to "let the data speak," as well as employing theoretical scrutiny, is one way of mitigating bringing a priori human biases into the model selection process. This approach also allows a wider set of models to be considered simultaneously, without ruling important ones out in advance. Hence, this enables new overlooked hypotheses to be considered and may offer a solution to the ambiguity documented above. Consequently, this is the key purpose of this study.

A NOVEL APPROACH THAT CONSIDERS OVERLOOKED, ALTERNATE HYPOTHESES

Under a bivariate problem framing that pits the "nature vs. nurture" inherited characteristics vs. experience-driven worldviews against each other, we would only investigate two hypotheses: VSWM => MT vs. MT => VSWM. However, in our triplet framing, the problem would now be more nuanced, involving an interim effect (see Figure 1). In traditional psychological research vocabulary, this is known as mediation (Baron & Kenny, 1986); in causal modeling vocabulary, discussed shortly, it is known as d-separation (Hayduk et al., 2003).

However, even as a triplet framing, to pit these two hypotheses against each other is still limiting and betrays a priori assumptions that there are only two worthy hypotheses to test. This would represent a less futile, but similar, error to that which Schellenberg (2020) warned against, since contemporary behavioral genetics research tells a story of nature *and* nurture, rather than nature *vs.* nurture (Plomin, 2018). As there is no necessary logical contradiction between the *inherited characteristics* and *experience-driven* hypotheses: both causal explanations could logically hold, and instead, it would be a matter of determining their degree

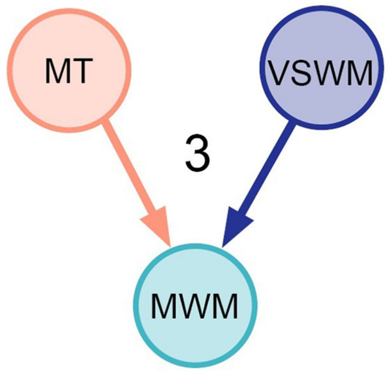

FIGURE 2. A novel hypothesis: the independent causes hypothesis (Model 3).

of relevance. In other words, a priori, it is logically plausible that concurrently A) nonmusical WM faculties, such as VSWM, support MWM, and B) MT is a factor that causally improves MWM faculties. Similar to mediation, a consequence of such a hypothesis is that the variables VSWM and MT need not be directly related to each other but can be related only via their relationship with MWM. We could call such a hypothesis, the "independent causes hypothesis" (model 3), which outlines the fact that MT and VSWM are only related, conditional on MWM, and both independently influence it (see Figure 2). However, note that, empirically, such a hypothesis is questioned by the documentation of positive correlations between WM and MT (Talamini et al., 2016, 2017), which should not be found under this causal structure. Nonetheless, model 3 represents one example of many such possible alternate models of the way MT, VSWM and MWM could interact. We spell these models out in Figure 4. and discuss some other alternatives later.

Model 3 and other similar models are of particular interest because of their "V" pattern. Fortunately, the pattern of correlational relationships that would underly this hypothesis, known in the causal modeling literature as a *V-structure* (Elwert, 2013), is one of the only mathematically proven correlational dependencies that can imply a causation (Pearl, 2009). Furthermore, if found in real data, it would suggest that there is no contradiction between the experience-driven and prior dispositions hypotheses. The next section of this paper presents the mentioned concepts of d-separation and V-structures in more detail to help the unfamiliar reader understand the causal modeling framework, and hence, the design of our study.

A MINI CAUSAL MODELING PRIMER FOR PSYCHOLOGICAL RESEARCH

Like structural equation modeling (SEM), Pearl's causal modeling theory (2000, 2009) employs graph theory to represent causal relationships visually but expands on traditional SEM terminology and concepts. In causal modeling theory, and graph theory more generally, a graph that hypothesizes causal relationships between variables and contains no feedback loop between them is called a directed acyclic graph (DAG; see Shrier & Platt, 2008, for a biomedical application). Figure 4 shows the possible a priori hypotheses represented as DAGs. Next, our analysis requires briefly outlining two concepts that are potentially less well known to the general psychological community: *d-separation* and collider/*V-structures* (for an extended but still very accessible introduction see Elwert, 2013).

D-separation is effectively equivalent to partial correlation, asking whether two variables (X and Y) become statistically independent given another variable or set of variables (Z). If X and Y become unrelated given Z, they are said to be d-separated. As suggested by the term d-separation, which stands for directional-separation, this has a causal implication: if the relationship of X and Y can be accounted for by Z, then there is no direct causal effect of X on Y, allowing any direct causal path between X and Y to be removed (Hayduk et al., 2003). Under this circumstance, Z is said to fully mediate the effect of X and Y and this is effectively equivalent to the traditional psychological methodological notion of mediation (Baron & Kenny, 1986).

The second causal tool we employ is a search for *V-structures* (or "colliders"). Pearl (2017) calls V-structures a "gift from God" because they imply causality. A V-structure occurs when two uncorrelated variables become correlated after adjusting for a third variable (Pearl et al., 2016; Tian & Pearl, 2013). Figure 3 visually displays the pattern of bivariate and partial correlations required to produce a V-structure: A) the bivariate correlations suggest X and Z are not significantly related, but X and Y are significantly related, and Y and Z are significantly related; B) the partial correlations find X and Z to be related conditional on Y. The pattern of bivariate and partial correlations in A and B respectively imply that X and Z cause Y but are unrelated with each other.[1] It is worth emphasizing that it is only the emergence of this special class of dependencies that would identify a unique causal solution. In other

[1] Classical suppression in a regression produces a similar pattern of partial regression coefficients, except with X and Y not being significantly correlated (e.g., Lewis & Escobar, 1986).

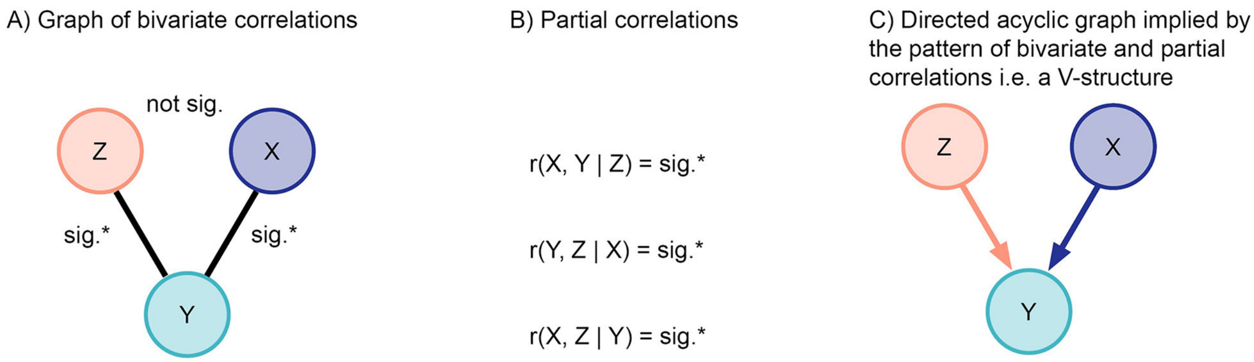

FIGURE 3. Graph showing the pattern of bivariate and partial correlations required to produce a V-structure.

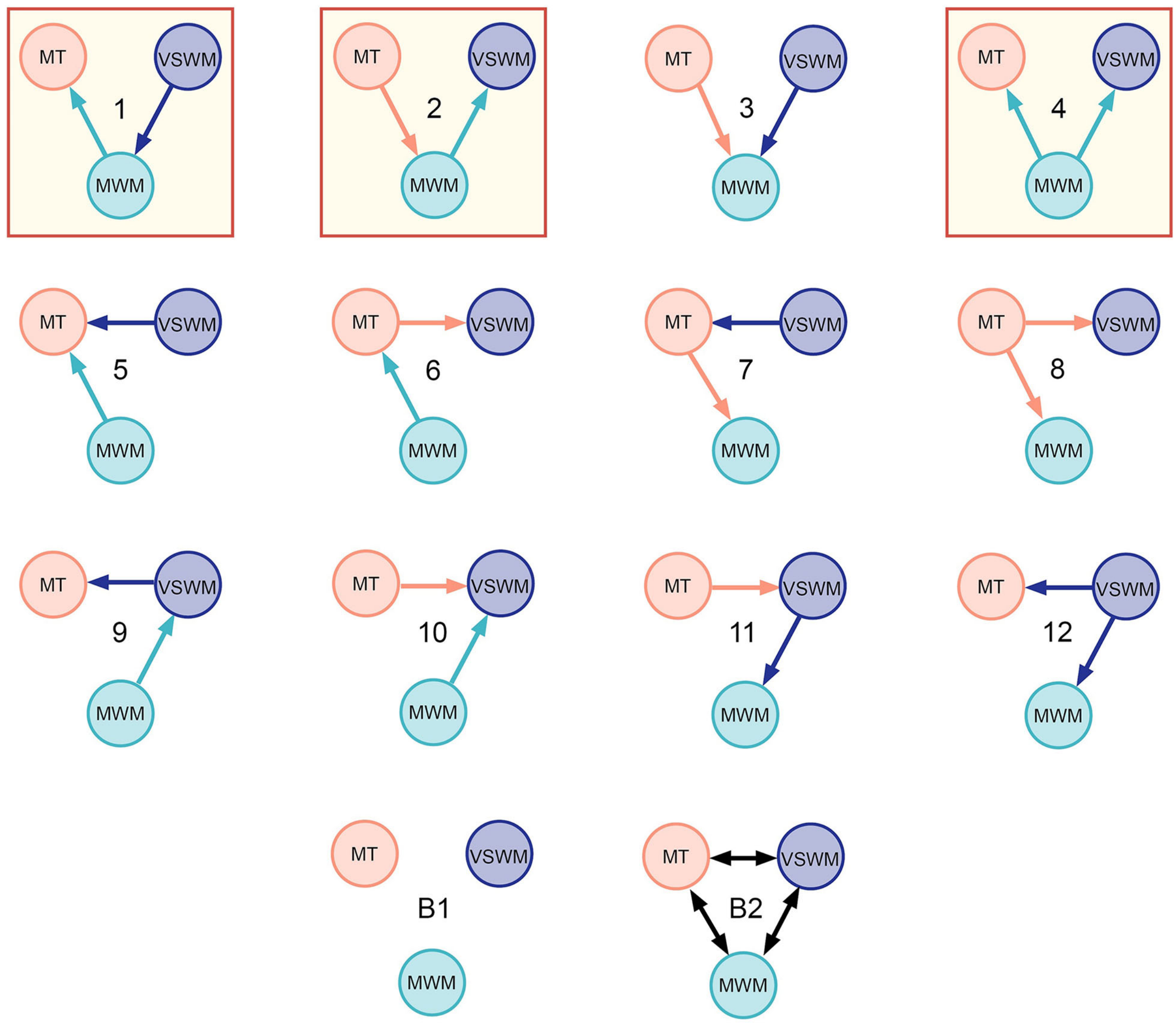

FIGURE 4. All possible directed acyclic graphs representing causal relationships between visuospatial working memory (VSWM), musical working memory (MWM), and music training (MT) where all factors are related causally in some form. The models which are eventually selected are highlighted.

words, other non-V-structure causal scenarios such as models 1 and 2 (see Figure 1) cannot be detected by this method.

To understand why this is so, consider an intuitive and musical example that helps to grasp the logic of a V-structure. Two factors that often account for enrollment in a school for gifted children are either A) musical talent or B) sports talent. In the general population, musical talent and sports talent are unrelated. However, within the school, musical talent and sports talent would be negatively correlated. In other words, if you know a person is a good hockey player and enrolled at the school, you can be fairly sure that her musical talent will be comparatively low! Hence, musical talent and sports talent become related conditional on enrollment status and one "explains away" the other (Pearl & Mackenzie, 2018). Assuming there are no confounding variables, finding empirically that musical talent and sports talent become related conditional on enrollment in a gifted school would respectively imply that sports talent and musical talent *cause* school enrollment and rule out other possible causal trajectories between the three variables: other causal patterns would give rise to another set of correlational dependencies.

In summary, in an empirical dataset, from the pattern of partial correlations, d-separation can tell us which two variables are directly related to each other after taking the influence of other variables into account. Additionally, from the pattern of bivariate and partial correlations we might be able to identify V-structures in triples of variables that would allow us to assign directional paths to the relationships between variables.

METHODOLOGICAL BENEFITS AND LIMITATIONS TO THE CAUSAL MODELING APPROACH

It is worth briefly considering some benefits of the causal modeling approach as well as its limitations. First, as mentioned, an issue with the approach that considers only models 1 vs. 2 is that it a priori rules out many other possible candidates and hence implies a strong assumption with a comparatively weak foundation. The ability to investigate other causal models, such as model 3 discussed earlier, was a key motivation for this research. But not only does the causal modeling approach allow us to consider this single overlooked hypothesis, it widens the hypothesis space of our question from the 2 directional hypotheses assumed in previous approaches to twelve by not ruling out any other models a priori. This gives all the same ability to be discovered in real data (see Figure 4). For instance, there are two other V-structures that are possible under our problem framing, only these seem marginally less likely. However, simply because we consider them less likely, our approach does not rule them out in the first place. Hence, due to the consequence of this method widening the hypothesis space and becoming more data-driven in nature, it may offer a novel solution to the ambiguous causal problem. Furthermore, while all the models enumerated in Figure 4 could be tested against each other using measures of global model fit (i.e., structural equation modeling), constraint-based causal modeling rules out many candidates beforehand which safeguards against assessing the fit of models that are causally implausible.

Second, since MT is regarded as the formal pursuit of acquiring musical skills, which is often intense and undertaken over longer time periods (Lehmann et al., 2007), potential transfer effects are generally thought to take place over months or years (Hyde et al., 2009). Consequently, employing experimental or quasi-experimental methods or even randomized control trials (the "gold standard" of making causal claims; Jones & Podolsky, 2015) to establish causal relationships between MT and other abilities becomes complex and expensive and such research is relatively scarce. Hence, in lieu of more longitudinal research, replicated in multiple settings, causal modeling would provide some tentative suggestions and potentially allow us to diagnose methodological issues (e.g., overlooking hypotheses a priori) before collecting data longitudinally.

A key limitation with the approach is the fact that, as Pearl points out, "data [alone] are profoundly dumb" (Pearl & Mackenzie, 2018, p. 6). Hence, data need to be supplemented by causal assumptions derived from substantial theory available in a research domain. Consequently, the data-driven approach is no panacea, and it is important to subject any statistically plausible models to theoretical scrutiny to avoid being theory-blind. Furthermore, the models that result from causal analysis should eventually be confirmed by longitudinal studies; though as noted, causal analysis may allow issues to be identified, as well as novel plausible hypotheses to be generated, in advance of longitudinal data collection.

THE PRESENT STUDY

The main purpose of this study is to reinvestigate the plausibility of previously overlooked hypotheses regarding the relationship between MT and nonmusical variables, such as VSWM, reframed as a triplet of variables that includes tests of musical memory. We hence seek to rule out ways the variables MT, VSWM, and MWM could interact and narrow down our wide a priori model set. In the best case scenario of finding a V-structure, we may be able to reason for a unique causal solution that may help solve the ambiguities discussed.

To attempt to identify the most plausible causal relationships between VSWM, MWM, and MT, we used a latent variable approach. For similar approaches see Visser et al. (2006) and Okada and Slevc (2018). The use of causal discovery tools to identify the possible causal directions between latent variables from cross-sectional data, and to discard causal relationships that do not have empirical support from the data, is still fairly novel in the analysis of behavioral psychological data. For initial work in this area see Moffa et al. (2017) and McNally (2016).

In our approach, first, we seek to validate a measurement model consisting of two sets of manifest variables thought to measure the latent variables VSWM and MWM; as well as a single measured variable to represent the latent variable MT. Then, using causal modeling, we attempt to identify a single causal model which explains the relationships between these variables based on the pattern of bivariate and partial correlations in our cross-sectional dataset.

HYPOTHESES

In Figure 4, we outline all the possible DAGs as hypotheses where each of the factors is related to at least one other factor in a directional manner. There are 12 such possible directed acyclic models which could explain the causal relationships between VSWM, MWM, and MT. Since the literature already suggests that MWM, VSWM, and MT are significantly related to one another (Anaya et al., 2017; Suárez et al., 2016; Talamini et al., 2016), we intend to go further and obtain a more precise *causal* explanation of the relationships. Hence, model B1 represents a hard null model, with no relationships between the factors at all, and model B2 represents a soft null model where the factors are all related as previously observed, but nondirectionally. All other models represent alternative hypotheses which are not ruled out a priori.

Models 1, 2, and 3 (the prior dispositions, the experience driven, and the independent causes hypotheses, respectively) have already been discussed. For the sake of brevity, we cannot discuss all the theoretical positions corresponding to the other nine models. Consequently, other than the three positions discussed, we limit ourselves to discussing one more model, model 4, and its related models, which we deemed implausible a priori.

Any model, such as model 4, which holds that MWM precedes VSWM causally seems unlikely. In terms of cognitive evolution, it is implausible that a specialized system would precede or support a more general one (Overmann et al., 2013), i.e., that individuals develop a specialized psychological system (e.g., MWM for handling musical symbols) which precedes the development of a general system (e.g., VSWM for generally handling visual information). Moreover, while some evolutionary psychologists argue for the importance of musical development in humans (Brown, 2017; Tarr et al., 2014), there is no compelling reason to suggest that musical abilities were more important to survival and reproduction than basic faculties such as VSWM (Overmann et al., 2013). Hence, these models, and some others, seem implausible, but we nonetheless allow them to be subjected to statistical scrutiny, with the condition that we discuss any remaining model's meaning and theoretical plausibility.

Other than model 3, note that, since they are V-structures, models 5 and 10 could also offer a unique causal solution that is detectable by our method, only we deem them to be less theoretically plausible. Furthermore, models that are unidirectional (e.g., models 1, 2, 6, etc.) cannot be individually detected by causal modeling since they do not comprise a V-structure. Hence, to argue for a unique causal solution, one of models 3, 5 or 10 must hold true. Otherwise, there will be a narrowed down set of models, leaving causal ambiguity.

Method

To measure the three latent variables, we tested participants on a battery of tasks and hypothesized how these would load onto the three factors. This constitutes our measurement model, which is described after the tasks below.

PARTICIPANTS

We recruited 148 participants aged 18–50 (to avoid strong effects of cognitive decline of older participants; Salthouse, 2009), and heterogenous on MT, through social media and on-campus advertising at Goldsmiths' College, London, United Kingdom, and Macquarie University, Sydney, Australia, to complete the task battery described below. The resulting sample had a mean age of 26.4 ($SD = 7.7$) and consisted of 88 females and 56 males (the age and gender of four participants was missing for unknown reasons). Eleven participants received course credits in exchange for their participation. All other participants received a small monetary compensation. Recruitment was predominantly from among the Goldsmiths and Macquarie student communities but there was also a proportion of participants recruited from off-campus. The study was granted ethical consent by both Goldsmiths' and Macquarie University ethical approval bodies and participants were free to opt out at any time. See the Appendix for descriptive statistics.

WORKING MEMORY (WM) TASKS

The construct of WM is defined as the ability to simultaneously store and actively transform information (Baddeley & Hitch, 1974). It is primarily concerned with transient and short-term phenomena as well as projection across time (Fuster, 1997). In this sense, WM is usually described as a more dynamical construct (Simmering & Perone, 2013), which is dissociable from the otherwise static-natured construct of intelligence (Alloway & Alloway, 2010). It is even argued to be the limited-capacity system that underlies and constrains intelligence, and hence has been referred to as an ultimate *cognitive primitive* (Alloway & Alloway, 2013). Since the following six tests (or "tasks" as they are often known as in the WM literature), all reflect the dynamical and transient storage and manipulation of stimuli, we operationalize them as measuring WM.

All of our tasks yield item response theory (IRT) scores, which are the main output of modern test theory (Embretson & Reise, 2000).

VSWM TASKS

VSWM tasks are designed to measure the visuospatial scratchpad element of Baddeley's (2000) WM model. In our three VSWM tasks, Jack and Jill (JaJ), Memory Updating Figural (MUF), and Backwards Digit Span (BDS), the common uniting element is a strong emphasis on transiently remembering and manipulating something presented in the visual domain.

Jack and Jill (JaJ)

JaJ (Silas et al., 2022) measures VSWM capacity based on a dual task paradigm and is similar to earlier versions of visuospatial dual task paradigms (e.g., Alloway, Gathercole, Kirkwood, & Elliott, 2008; Shah & Miyake, 1996). Participants must hold multiple spatial locations on a hexagon in WM while answering an unrelated question for each location point shown. Two characters, Jack and Jill, are presented, both holding a ball in one of their hands. For each image, participants have to: 1) decide whether Jack holds the ball in the same hand as Jill, and 2) remember the position of Jack's ball on the hexagon of dots. At the end of each sequence, participants must indicate the position of the balls in the correct order. The task had 14 trials with the length of sequences increasing and hence becoming more difficult. IRT scores for the JaJ task were generated online using the R package psychTestR v2.13.2 (Harrison, 2020) according to an underlying explanatory IRT model (Silas et al., 2022).

Memory Updating Figural (MUF)

The MUF task is a VSWM task similar to the Salthouse et al. (1991) task and is, more specifically, a reimplementation Vock and Holling's (2008) VSWM task. Participants were presented a variable number of rectangles where dots could appear in any corner for 1.5 seconds at a time, followed by arrows pointing to other corners of the same rectangles. Participants had to remember the various dot locations, imagine where the dots would move to based upon the arrows shown and click in the corners of empty rectangles indicating the final position of each dot. The task comprised 14 items increasing in difficulty based on the number of mental operations to be completed. We could not use previous IRT models for this task because they were constructed using a different task battery and a sample of children, and therefore would likely not produce realistic scores in the context of our analysis. Consequently, we computed post hoc IRT scores for each participant based on sum scores and using the Rasch psychometric model. A Pearson's product-moment correlation of the sum scores with the IRT scores yielded a strong correlation ($r = .98$, $p < .01$), indicating a high level of consistency between the classical test theory and IRT scoring methods.

Backwards Digit Span (BDS)

BDS tasks represent a classic measure of WM (Case & Globerson, 1974). Participants must remember an ordered sequence of digits, mentally reverse it, and enter the reversed sequence by clicking the numbers on a keypad. We reimplemented the version used by Vock and Holling (2008), which consisted of 12 trials of increasing difficulty using sequences of four to seven digits length. Since all stimuli were presented in the visual domain and responding involved clicking digits on a visually displayed keypad that spatially organized the digits, we consider this a visuospatial BDS task. Previous item response theory (IRT) models for this task could not be used as they were constructed based on their inclusion in another battery of tasks (as explained above). We computed IRT scores for each participant by fitting a Rasch psychometric model to the collected data. A Pearson's product-moment correlation of the sum scores with the IRT scores yielded a very strong correlation ($r = .99$, $p < .01$), indicating a high level of consistency between the classical test theory and IRT scoring methods.

MWM TASKS

MWM tasks are designed to measure music-specific WM abilities that have been argued to not be sufficiently explained by the widely accepted Baddeley and Broadbent (1983) model (e.g., see Berz, 1995). Since many people do not engage in developing musical expertise (unlike VSWM, which any sighted person, and perhaps nonsighted persons too, must naturally develop), it has been argued that such WM systems represent a different class of WM called long-term working memory (Ericsson & Kintsch, 1995), which particularly relies on experience-driven and domain-specific training.

Music-specific ability tests should concern musical, rather than simply auditory, stimuli. Musical features are more sophisticated than auditory, involving a knowledge of syntax and structure (e.g., in the realms of rhythm, harmony, and melody). For example, our Melodic Discrimination Test (MDT) uses pastiche Irish folksong melodies, and so knowledge of the diatonic scale and Western melodic transition probabilities should help memory. In our three MWM tasks, Rhythm Ability Test (RAT), Melodic Discrimination Test (MDT), and Pitch Image Arrow Task (PIAT), the common uniting element is a strong emphasis on transiently remembering and manipulating phenomena which are musical (with melodic, harmonic, or rhythmic qualities).

Rhythm Ability Test (RAT)

The RAT (Müllensiefen et al., 2019) measures memory for non-pitched rhythmic stimuli and is related to the musical sequence transcription task described by Zuk et al., (2013). Each trial plays a rhythmic pattern of high frequency claps and low frequency bass drum kicks. Afterwards, visual representations of four different rhythms are shown with light blue squares representing claps on a top row and dark blue squares representing bass drum kicks on a bottom row. Participants must click the visual representation that corresponds to the rhythmic pattern they heard. The test comprised 16 trials of increasing difficulty as a function of number of rhythmic events, the complexity of the rhythm, and the similarity of the target sequence to the three lures. IRT scores for the RAT task were generated using the R package *psychTestRCAT* v1.0.2 (Harrison, 2018) according to an underlying explanatory IRT model (MacGregor, Müllensiefen, Fiedler, Andrade, Forth, & Frieler, 2022). The RAT can be considered a WM task because it requires transiently holding information in memory and mentally transforming stimuli from musical to visual modalities.

Melodic Discrimination Test (MDT)

We assessed melodic discrimination ability using the adaptive MDT (Harrison et al., 2017). This test uses a 3-AFC response paradigm with each item consisting of three versions of a melody played at different transpositions in pitch (for example: first: D major, second: Eb major, third: E major). Two of these versions are always identical and one is always different. The participant's task is to identify the nonidentical melody, but to ignore transpositions between versions. The version of the MDT used in this study comprised 20 items using an adaptive procedure (Harrison et al., 2017). IRT scores for the MDT task were generated online using the R package *psychTestRCAT* v1.0.2 (Harrison, 2018) according to the underlying IRT model described in Harrison et al. (2017). Based on the process model described in Harrison et al., (2017), the MDT can be considered a WM task.

Pitch Imagery Arrow Task (PIAT)

The PIAT has been established as a valid and reliable measure of musical imagery, the ability to mentally represent and transform pitch (Gelding et al., 2020). Participants hear a tonal context comprising of an ascending major scale followed by the tonic of the scale. Starting from the tonic or dominant, a series of notes are played going either up or down one note in the major scale. Arrows that match the direction (up or down) of the change in tonal sequence are presented simultaneously. After a variable number of times, the audio cue ceases, leaving just the presentation of the arrows. The participant must imagine a continued progression along the scale as guided by the arrows, but with no further audible reference. At the end of a trial, a probe tone is played. The participant indicates whether this tone matched the tone they were imagining. A correct response requires identifying the correct place to end up in the scale based on the arrow indications. The task was adaptive. IRT scores for the PIAT task were generated online using the R package *psychTestRCAT* v1.0.2 (Harrison, 2018) according to the underlying explanatory IRT model (Gelding et al., 2020).

MT Measure (Gold-MSI)

MT involves the development of skills such as performing, memorizing, composing, sight-reading, and the aural identification of music (Lehmann et al., 2007). People who have developed expertise in music are often

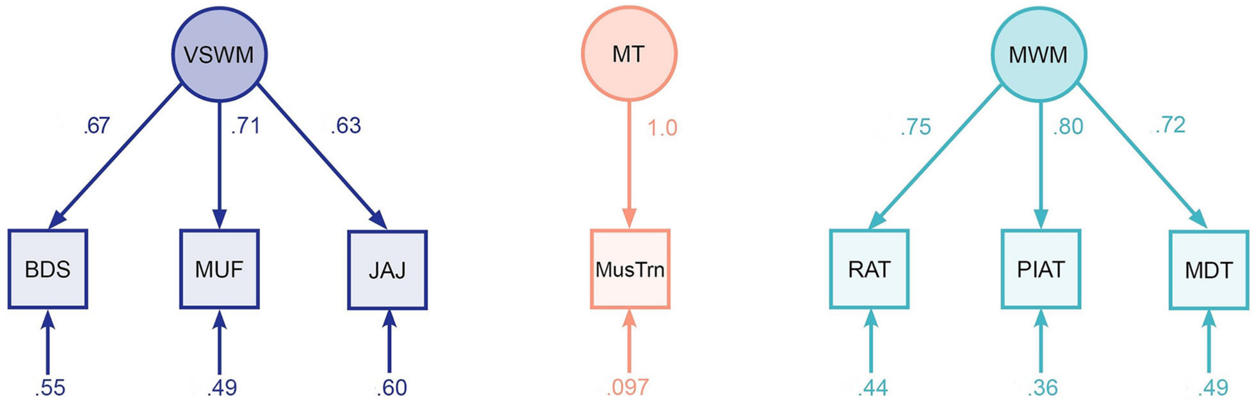

FIGURE 5. Measurement model for the VSWM (visuospatial working memory), musical working memory (MWM), and music training (MT) latent variables. Latent variables are shown in circles and measured indicators in squares. Arrows pointing from latent variables to measured indicators represent factor loadings and arrows pointing up towards measured indicators represent measurement error values.

referred to as "musicians" in the literature, although this is a somewhat artificial construct, as MT can be measured on a continuous scale without dichotomization (Müllensiefen et al., 2014). The MT latent variable was defined using the score yielded from the Gold-MSI's 7-item MT subscale as the single manifest indicator of latent MT, together with its measurement error (derived from its internal reliability as estimated in Müllensiefen et al., 2014). Defining a latent construct by a single indicator has the advantage that the known measurement error can be considered. For more on the practice of single indicator variables see Hayduk and Littvay (2012).

MEASUREMENT MODEL

We specified a measurement model with three VSWM tasks (Jack and Jill, JAJ; Memory Updating Figural, MUF; Backwards Digit Span, BDS) to load onto the VSWM factor and three MWM tasks (Rhythm Ability Test, RAT; Melodic Discrimination Test, MDT; Pitch Imagery Arrow Task, PIAT) to load onto the MWM factor. MT was measured using the MT subscale of the Gold-MSI self-report inventory. Its measurement error was specified based on the empirical reliability ($\alpha = .903$) as reported in Müllensiefen et al. (2014). All measures produced continuous scores. The MUF and Gold-MSI tasks were implemented in Qualtrics (Qualtrics, 2018) whereas the BDS, JAJ, RAT, PIAT, and MDT tasks were implemented using the R package *psychTestR* v2.13.2 (Harrison, 2020). Figure 5 shows a diagram of the measurement model with the empirical loadings derived from the subsequent analyses.

PROCEDURE

All participants completed the task battery under quiet laboratory conditions in individual test cubicles. Each task was presented inside of an internet browser and had an introduction as well as example and/or training trials. A researcher was available at all times to answer questions or help with technical difficulties. The order of tasks taken was identical for all participants: BDS, MUF, JaJ, RAT, MDT, PIAT, Gold-MSI. Participants required an average of approximately 60 minutes to complete the full battery.

DATA ANALYSIS

Our analysis consists of three fundamental steps:

1. Assess the measurement models of the VSWM and MWM latent variables using exploratory factor analysis and generate factor scores.
2. Assess the pattern of correlations and partial correlations between the MWM, VSWM, and MT factors to: A) apply the rules of d-separation and evaluate whether edges can be removed from a fully connected causal graph, and B) assess whether V-structures emerge. This process should whittle down the candidate model set (Figure 4) to a smaller number of models but hopefully identify a single V-structure model.
3. A) Fit a SEM (which accounts for the measurement model and the relationships between the latent variables simultaneously) to the observed data based on the results of step 2.
   B) Assess global model fit and statistical significance.

TABLE 1. *Pairwise Correlations of All Tasks*

			VSWM			MWM		
		MT	**BDS**	**MUF**	**JaJ**	**RAT**	**PIAT**	**MDT**
	MT		.28**	.31**	.23*	.46***	.57***	.52***
	BDS			.53***	.38***	.40***	.42***	.30**
VSWM	**MUF**				.47***	.46***	.42***	.29**
	JaJ					.53***	.46***	.32**
	RAT						.54***	.58***
MWM	**PIAT**							.56***
	MDT							

Note: Significance is denoted as *p < .05 **p < .01 ***p < .001

TABLE 2. *Exploratory Factor Analysis Results for the VSWM and MWM Factors*

VSWM				MWM			
Task	**Factor loading**	**h^2**	**u^2**	**Task**	**Factor loading**	**h^2**	**u^2**
BDS	.66	.43	.57	**RAT**	.75	.56	.44
MUF	.80	.64	.36	**PIAT**	.72	.52	.48
JaJ	.58	.34	.66	**MDT**	.78	.61	.39

Results

BIVARIATE CORRELATIONS

First, we assessed the correlations between all measures (see Table 1). All scores correlated significantly (all p values < .05 after correcting for multiple comparisons using Holm's procedure). The correlations between the VSWM tasks were moderate ($r = .38$ to .53), as were the correlations between the MWM tasks ($r = .54$ to .58). MT showed moderate correlations with the MWM tasks ($r = .46$ to .57), and small to moderate correlations with the VSWM tasks ($r = .23$ to .31). The VSWM and MWM tasks had small to moderate correlations with each other ($r = .29$ to .53).

FACTOR ANALYSES

The hypothesized measurement models for the VSWM and MWM factors were assessed with two independent minimum residual exploratory factor analyses. All factor loadings were > .50 for the VSWM factor and > .70 for the MWM factor, which indicated that the tasks/items represented the factors well (see Table 2). The VSWM latent variable explained 47% of the variance in the observed VSWM task scores while the MWM latent variable explained 56% of the variance in the observed MWM task scores. Factor scores were extracted for each participant on each variable.

We then assessed the pairwise relationships of the extracted VSWM, MWM, and MT factor scores using Pearson's correlation coefficient (see Table 3). All factors correlated significantly with each other (all p values < .001 after correcting for multiple comparisons using Holm's procedure). The relationships between MT and MWM ($r = .63$) and MWM and VSWM ($r = .54$) were substantial and the relationship between MT and VSWM ($r = .36$) was moderate. Consequently, on the basis of pairwise correlations alone, null model B1, which posits no correlations between the factors, can be rejected.[2]

TABLE 3. *Bivariate Correlations of VSWM, MWM and MT Factors*

	VSWM	**MWM**	**MT**
VSWM		.54***	.36***
MWM			.63***

Note: Significance is denoted as *p < .05 **p < .01 ***p < .001

As outlined above, the comparison of correlations and partial correlations allows us to manually assess d-separation and V-structures. The pattern of correlations and partial correlations between the VSWM, MWM, and MT factors reveals that, while all three factors are significantly related as bivariate

[2] Despite the cross-correlations between VSWM and MWM after extraction via our method of two independent factor analyses, submitting all six variables to the same parallel and factor analyses nonetheless suggested two factors. Under this alternative analysis, the hypothesized measures loaded on to the hypothesized factors and no substantial cross-loadings emerged (absolute values ranging .05 to .26 cross loading onto the other factor).

TABLE 4. *Partial Correlations of VSWM, MWM and MT Factors*

	Pearson's *r*
r(VSWM, MWM | MT)	.42**
r(VSWM, MT | MWM)	.05
r(MWM, MT | VSWM)	.54**

Note: Significance is denoted as *p < .05 **p < .01 ***p < .001

correlations, conditional on MWM, MT, and VSWM become unrelated; see Table 4 (whether correcting for multiple comparisons using Holm's, procedure or not). According to the rules of d-separation, this suggests that there is no direct effect of MT on VSWM or vice versa. Consequently, in a directed acyclic graph that shows all three factors as related by edges connecting them, the edge connecting MT and VSWM can be removed. This creates a scenario whereby MWM mediates the relationship between VSWM and MT. This process of d-separation allows us to reject nine models from Figure 6 as well as the null models. However, the pattern of dependencies for a V-structure described earlier—that two uncorrelated variables become correlated after adjusting for a third variable—does not occur here. Hence, model 3 (and other models) is precluded based on the empirical implications. As a result of the pattern of dependencies, only models 1, 2, 4 remain plausible. All other models imply different dependencies to the observed pattern of bivariate and partial correlations.

Consequently, since no V-structures emerge, no unique directed causal relationships among the variables are suggested by the pattern of bivariate and partial correlations. The PC-algorithm (Spirtes et al., 2000) implements the rules of d-separation and a search for V-structures, allowing for a computational search for causal relationships from correlational data. A fully automated analysis using the R package *PCalg* (Kalisch et al., 2012), which implements the PC-algorithm, confirmed that only models 1, 2, and 4 are supported by the data.

STRUCTURAL EQUATION MODELING

Three independent SEMs representing models 1, 2, and 4 were fit to the data of the full sample ($N = 148$) using the function sem from the R package lavaan v0.6.9 and the robust full-information maximum likelihood method (Rosseel, 2012). See Table 5 for an overview of the results. For all models, two commonly used indices indicated a satisfactory model fit in absolute terms (SRMR = .05, CFI = .95; Kline, 2012). However, two other commonly used absolute fit indices suggested a less than ideal, but still acceptable, model fit: RMSEA = .10, TLI = .92; $\chi^2(13) = 29.39$, $p < .01$. The Bayesian Information Criterion (BIC) as an indicator of relative model fit and the results of a chi-squared test both suggested that models 1 and 2 fit the data significantly better than the null B1 model of no correlations between the factors, ΔBIC = 108.8; $\chi^2(15) = 147.98$, $p < .001$. However, the null B2 model of full bivariate correlations between all three factors had a BIC value that was only slightly worse than the BIC of models 1 and 2 (ΔBIC = 1.8). The chi-squared test comparing model fits also suggested a near-equivalent fit of the null B2 model by narrowly missing significance on a likelihood ratio test of difference in model fit, $\chi^2(12) = 26.32$, $p = .06$. Figure 6 shows models 1, 2, and 4 on the VSWM, MWM, and MT latent variables with standardized factor loadings (single-headed arrows) pointing from the latent variables to each other and parameter estimates representing the magnitude of association (see Figure 5 for the empirical measurement model loadings).

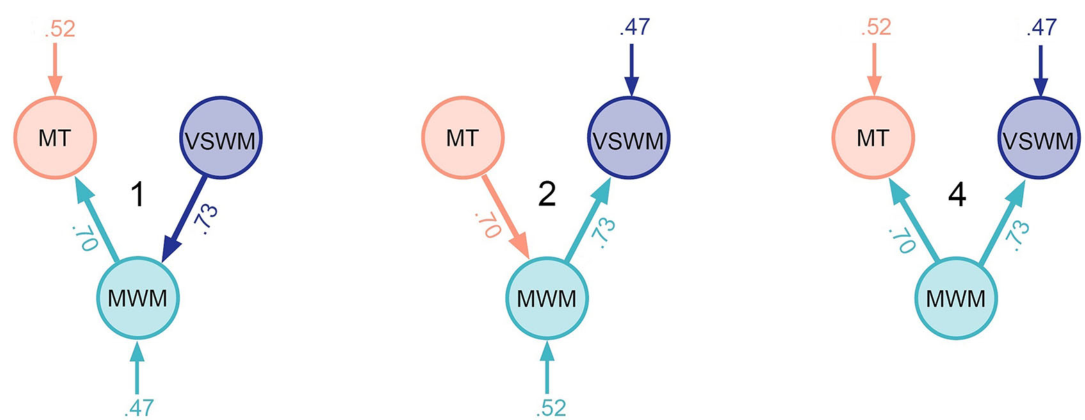

FIGURE 6. Statistically plausible SEM path diagrams (models 1, 2, and 4).

TABLE 5. *Relative Model Fit Indices*

Model(s)	No. of estimated parameters	*df*	*df* diff	AIC	BIC	Chi square	Chi-square diff	*p value*
Null B1	13	15		3235.6	3273.3	147.98		
1, 2 & 4[†]	15	13	1	3121.1	3164.5	29.39	117.40	< .001
Null B2	16	12	3	3120	3166.3	26.32	3.67	.06

[†] Note that, as explained in text, due to coming from the same model set, models 1, 2 and 4 have equivalence to one another from the perspective of a SEM. Consequently, their fit indices will be the same, and they cannot be compared with one another.

However, since models 1, 2, and 4 have model equivalence (Williams, 2012) with each other, they have indistinguishable statistical implications in an SEM framework. Therefore, model testing cannot tell us which of the three remaining models is more plausible as they produce the same parameter estimates and model fits despite being defined with opposing causal directions. However, good models of global fit via SEM adds validity to the results of causal modeling since it suggests the models do not contradict the observed data and are hence plausible.

Discussion

This study sought to explain the relationships between MT and various WM constructs. We suggested that the noninclusion of musical memory tests in previous similar studies was an important limitation that we argued may produce a more ambiguous view of such relationships. To rectify this, we specifically sought to empirically characterize MWM and VSWM faculties through latent variable modeling and then assess how these factors are related to MT. By framing the problem as a triplet of variables, we opened the possibility of discovering both mediation (alternatively known as d-separation) and V-structures, the latter of which would suggest an unambiguous causal solution to question. We also noted that there was no necessary logical contradiction between inherited characteristics and experience-driven approaches and that were several other plausible causal models that could explain the relationships between MT, MWM, and VSWM. Previously, such models may have been overlooked (e.g., by a nature *vs.* nurture worldview). Consequently, our approach challenged having strong a priori causal assumptions and tried to avoid the misinterpetive biases often seen when studying MT and nonmusical cognitive abilities (Schellenberg, 2020).

As Schellenberg (2020) noted, these biases often involve incorrectly inferring causality from cross-sectional data. However, we showed that it is possible to legitimately infer causality from correlational data under a very strict set of conditions. More broadly, our approach involved the more traditional psychometric approach of latent variable modeling. Consequently, first, we had to assess our measurement model. The factor analyses confirmed the validity of the hypothesized measurement models for the VSWM and MWM constructs as presented in Figure 5. Considering that the combination of these tasks was novel, this is an achievement which suggests that future research (e.g., longitudinal) could similarly combine the tasks.

In accordance with findings from the existing literature (e.g., Hambrick et al., 2018; Talamini et al., 2017) the VSWM, MWM, and MT latent variables all had significant, positive, and moderate to substantial bivariate relationships with one another. From the a priori space of 12 models that the causal modeling approach opened up (Figure 4), the process of d-separation allowed us to narrow our model set down to three statistically plausible causal models. All three remaining models suggest that there is no direct relationship between VSWM and MT but that a domain-specific MWM mediates the relationship. However, there was no V-structure in the pattern of dependencies, meaning that it was not possible to identify a single causal model, as we had hoped for.

Nonetheless, it can still be considered a substantial insight to narrow down the set of candidate models by discarding nine models, as well as the null models, which were included a priori but not in accordance with the implications of the empirical data. Traditional SEM confirmed the plausibility of our results through acceptable measures of global model fit. However, due the equivalence problem, and the fact that a SEM cannot itself discover causal patterns in data (Visser et al., 2006), we were not able to further decide between the three remaining candidate models in a data-driven manner.

Of the remaining models, model 4 was flagged as theoretically implausible a priori because it seems unlikely that a specialized system (MWM) logically precedes a related more general system (VSWM). Therefore, we rule this model out on theoretical grounds, and, in sum, our data suggests that models 1 and 2 represent the most likely situation of the models shown in Figure 4.

NO FAR TRANSFER WITHOUT NEAR TRANSFER

D-separation (similar to mediation) suggested that VSWM and MT were found to be unrelated, conditional on MWM, which suggests that VSWM and MT are not directly related to each other. However, while this does not imply a V-structure pattern, which would require VSWM and MT to be unrelated as a bivariate correlation but become related as a partial correlation (conditional on MWM), it implies that if there are any effects of MT and VSWM on one another, MWM mediates that effect. This represents an important finding because it suggests that there is *no far transfer without near transfer*: MT and VSWM are only related *via* MWM, which is known as a mediator effect (Baron & Kenny, 1986).

By implication, tests of musical memory should be included when studying the effects of A) MT and WM on one another and B) cognitive abilities and prior dispositions on musical engagement and training. This has important practical significance because many studies that measure the effect of music interventions on cognitive abilities do not include any tests of musical abilities and hence could be missing something important (e.g., Bailey & Penhune, 2012; Diaz Abrahan et al., 2019; Schellenberg, 2004). This demonstrates how latent variable methods used with cross-sectional data, as employed in this study, can help to inform the design of expensive longitudinal research in advance by flagging such issues. Along the same lines, in principle, from the wider a priori space opened by the causal modeling approach, a surprising but reasonable causal pattern could have been detected in the data. This would have allowed an intervention or longitudinal study to be designed in the appropriate way to capture an effect.

ALTERNATIVE EXPLANATIONS

The possible reasons for our method not to detect a clear causal structure suggest several alternative theoretical explanations as well as some methodological limitations. First, our analyses are conducted at the group level which assumes that effects are systematic/nomothetic. However, perhaps causal trajectories are better explained at the individual/idiographic level (see Hommel & Colzato, 2017) for a discussion of the approaches in psychology).

Second, the selection of VSWM and MWM, as variables to study with respect to MT, was principled but does not imply that if we had focused on other cognitive constructs, we would have obtained similar results. It is possible that other similar constructs may be sufficient to detect a unique causal effect with respect to MT. Other nonmusical constructs rather than VSWM, for example, could be related to language, mathematical, and sensorimotor abilities as well as social cognition. On this note: operationalizing variables related to WM is particularly challenging. The nature of WM as a cognitive primitive (Alloway & Alloway, 2013) suggests that *any* combination of WM tasks should contain common variance that represents some very general latent aspect of WM, like a central capacity constraint (Cowan, 2010; Vergauwe et al., 2010). Beyond this, the common variance extracted from a task set should reflect what is otherwise common about the combined tasks. In this vein, we reasoned judiciously what the common variance in each task set (VSWM, MWM) represented. However, this is nonetheless an interpretation.

Third, the true causal pattern could be unidirectional (e.g., either models 1 *or* 2) after all. However, causal modeling cannot detect this kind of pattern. Furthermore, this would still raise the question as to why there is evidence for both scenarios. Conversely, fourth, perhaps both causal trajectories described in models 1 and 2 are *simultaneously* true to some extent. This possibility is suggested by our null B2 model, with bivariate relationships connecting all variables, being very close to model 1 and 2 in terms of absolute model fit. Null model B2 implies that there are cyclic processes at work that cannot be investigated with the methods used here or that unmeasured variables influence both measured variables in a bivariate relationship.

In general, any unmeasured variables that mediate or moderate the relationships between MT, MWM, and VSWM may obscure our ability to detect effects between them. Socioeconomic status is an example of such a possible confounding variable. However, the broader debate about to what degree socioeconomic status and general life outcomes depend on predispositions goes beyond the scope of the current paper (see Plomin, 2018, for more on this). Perhaps more difficult is the presence of genetic variants which simultaenously affect two different traits. In studies of MT, a relevant illustration is provided by Mosing et al. (2014), who showed through modeling data from a large sample of twins that musical ability and musical practice may be a result of the same genotypic characteristics, an effect known as genetic pleiotropy. This is also a case of "the nature of nurture" (Plomin, 2018), whereby ostensibly environmental variance ends up ultimately being explained by genotypic variance. MT is, in theory and name, an environmental measure. However, it is very likely that a large proportion of its variance could be explained by genotypic factors. In a nongenetically sensitive design like ours, we cannot consider our MT variable as a pure measure of the environment. Hence, we do not really know to what extent our variables, MT, MWM, and VSWM are explained and driven by shared genotypic factors, which may blur any experience-driven effects. Moreover, we do not know how the relative contributions of nature and nuture interact. For a comprehensive discussion of the above issues, we recommend the reader Plomin (2018) and the empirical literature reviewed there.

Genetically insensitive designs require assumptions about (or ignorance of) relative genetic and environmental influences. Because such designs cannot disentangle these influences, this may explain why there is significant evidence for supposedly competing causal effects in studies of MT (e.g., Meinz & Hambrick, 2010; Ruthsatz et al., 2008). In other words, were genetic and environmental influences disentangled, an appropriate modeling strategy could make sense of the way in which the effects interact with each another, rather than assuming the underlying hypotheses compete. We remind the reader of our observation which drove this study, that there is no logical contradiction between models 1 and 2, but that they could mutually coexist in the same causal universe. Since our current modeling approach with cross-sectional data is relatively simplistic, it is only an approximation to the true underlying processes which might be better captured by gene-environment interaction models (Dickens & Flynn, 2001) or so-called multiplier models (Ceci et al., 2003; Dickens, 2007). A test of these models will require longitudinal data, but causal modeling results using our current dataset can still be interpreted to suggest the relative strength of the individual components.

A final methodological limitation to discuss involves latent variable methods, which are limited in psychological research in general. As mentioned, numerous other variables could affect the interactions between the MT, MWM and VSWM variables.[3] However, the prime reward of latent variable methods—yielding reduced (lower, minimized) measurement error through measuring the same constructs with multiple measures—also has the downside of increasing testing time multiplicatively, due to the number of measures required for each factor (it is typically desirable to have more than two indicators per variable, otherwise additional assumptions must be met; Beaujean, 2014). Consequently, latent variable approaches quickly bring testing constraints, which means many relevant variables must go unmeasured in psychological research. The axioms of causal modeling (Pearl, 2009) vehemently stress how important it is to stringently control for confounding variables to make causal inferences. If one could hold constant for such extraneous variables, it may be possible to achieve a purer picture of the interactions between the variables of interest. But to overcome this with cross-sectional data would require ingenuity, perhaps by combining our method with adaptive testing (Harrison & Müllensiefen, 2018) and planned missing designs (Graham et al., 2006).

AN ULTIMATELY HIERARCHICAL WORKING MEMORY SYSTEM?

Finally, we consider one final, important interpretation. Based on theory, we assumed VSWM and MWM to be distinct, domain-specific WM constructs. While different subconstructs of WM can be experimentally dissociated in terms of performance on different tasks (e.g., Alloway et al., 2006), it does not follow that they are entirely distinct constructs. For instance, previous research suggests that a tonal WM network in musicians has structural overlap with a verbal memory network (Schulze et al., 2011).

All our tasks demonstrated positive manifold, i.e., all positive correlations with one another. This reinvokes the widely replicated observation that there is component variance common to all tests of ability, known as a *g-factor* (Colom et al., 2004; Spearman, 1904). It seems reasonable to posit that there is ultimately one unitary WM (or indeed, cognitive) system that encompasses both VSWM and MWM constructs as components and that such a system is better modelled hierarchically with a single enveloping factor. However, it is beyond the scope of this study to test this.

The concept of a hierarchical system weakens the notion of transfer effects between "disparate" constructs. Instead, it would suggest a more hierarchical phenomenon where, while some abilities are relatively disparate, they are nonetheless ultimately related by very general processes. The presence of a g-factor could be interpreted as supporting inherited characteristics approaches, in that it suggests there is an ultimate general ability that may predispose people towards cognitively demanding activities such as MT (Colom et al., 2004; Spearman, 1904). There is a reasonable amount of evidence to suggest that there are central (i.e., non-domain-specific) aspects to WM (Cowan, 2010; Vergauwe et al., 2010). Therefore, it is likely that constructs, as statistically dissociated by psychological tests, capture facets of a unitary WM system that can be delineated in various ways depending on the idiosyncratic

[3] On this note, causal modeling is not strictly necessary for a problem with only three variables. Instead, it is particularly efficient if the number of variables is large and, consequently, the true model is more complex than possible in a triplet of variables.

combination of tasks employed. Furthermore, it may be that WM itself is not a system (i.e., a set of components and dynamical relationships) as such, but actually a way of processing sequential information in the brain (Christiansen & Chater, 2016). This would imply that similar functional properties should be observed in all domains, from music, to chess.

In general, MT provides an interesting framework in which to research WM because MT particularly requires manipulating multimodal phenomena. Perhaps developing musical expertise encourages WM to create more efficient cross-modal representations, blurring supposedly distinct subsystems. In any case, the strong multimodal nature of MT seems to question the underlying meaning of a multicomponent view of WM. It speaks less to engaging in the statistical dissociation of categorical constructs, which may be somewhat arbitrarily dissociated, and places emphasis back on researching general processes: the ones that help us with all cognitive abilities whether it is driving a car, playing music from notation, or remembering a telephone number. Under this framework, we may better understand and model the complex nature of the cognitive system, where causal pathways do not compete, but interact.

Author Note

The study was granted ethical consent by both Goldsmiths, University of London, UK and Macquarie University, Sydney, Australia ethical approval bodies.

Sebastian Silas has been supported by a doctoral scholarship from the Deutscher Akademischer Austausch Dienst (DAAD; German Academic Exchange Service). Daniel Müllensiefen has been supported by an Anneliese Maier-Research Prize awarded by the Humboldt Foundation. Peter Harrison was supported by a doctoral studentship from the EPSRC and AHRC Centre for Doctoral Training in Media and Arts Technology (EP/L01632X/1). No potential competing interest was reported by the authors.

Correspondence concerning this article should be addressed to Sebastian Silas, Department of Musicology, Hochschule für Musik und Theater, Hanover, Neues Haus 1, 30175 Germany. E-mail: silass@stud.hmtm-hannover.de

References

Alloway, T., & Alloway, R. (2010). Investigating the predictive roles of working memory and IQ in academic attainment. *Journal of Experimental Child Psychology*, *106*, 20–29. https://doi.org/10.1016/j.jecp.2009.11.003

Alloway, T. P., & Alloway, R. G. (2013). *Working memory: The connected intelligence*. Psychology Press.

Alloway, T., Gathercole, S. E., Kirkwood, H., & Elliott, J. (2008). Evaluating the validity of the Automated Working Memory Assessment. *Educational Psychology*, *28*(7), 725–734. https://doi.org/10.1080/01443410802243828

Alloway, T. P., Gathercole, S. E., & Pickering, S. J. (2006). Verbal and visuospatial short-term and working memory in children: Are they separable? *Child Development*, *77*(6), 1698–1716. https://doi.org/10.1111/j.1467-8624.2006.00968.x

Anaya, E. M., Pisoni, D. B., & Kronenberger, W. G. (2017). Visual-spatial sequence learning and memory in trained musicians. *Psychology of Music*, *45*(1), 5–21. https://doi.org/10.1177/0305735616638942

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2

Baddeley, A. D., & Broadbent, D. E. (1983). Working memory. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *302*(1110), 311–324. https://doi.org/10.1098/rstb.1983.0057

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. 8, pp. 47–89). Academic Press. https://doi.org/10.1016/S0079-7421(08)60452-1

Bailey, J., & Penhune, V. B. (2012). A sensitive period for musical training: Contributions of age of onset and cognitive abilities. *Annals of the New York Academy of Sciences, 1252,* 163-170.

Bangert, M., & Altenmüller, E. O. (2003). Mapping perception to action in piano practice: A longitudinal DC-EEG study. *BMC Neuroscience*, *4*, 26. https://doi.org/10.1186/1471-2202-4-26

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182. https://doi.org/10.1037/0022-3514.51.6.1173

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

Bergman Nutley, S., Darki, F., & Klingberg, T. (2014). Music practice is associated with development of working memory during childhood and adolescence. *Frontiers in Human Neuroscience*, *7*, 926. https://doi.org/10.3389/fnhum.2013.00926

Berz, W. A. (1995). Working memory in music: A theoretical model. *Music Perception*, *12*(3), 353–364. https://doi.org/10.2307/40286188

Bigand, E., & Tillmann, B. (2021). Near and far transfer: Is music special? *Memory and Cognition, 50,* 339–347. https://doi.org/10.3758/s13421-021-01226-6

Brown, S. (2017). A joint prosodic origin of language and music. *Frontiers in Psychology, 8*. https://doi.org/10.3389/fpsyg.2017.01894

Case, R., & Globerson, T. (1974). Field independence and central computing space. *Child Development*, *45*(3), 772–778. JSTOR. https://doi.org/10.2307/1127844

Ceci, S. J., Barnett, S. M., & Kanaya, T. (2003). Developing childhood proclivities into adult competencies: The overlooked multiplier effect. In R. J. Sternberg & E. L. Grigorenko (Eds.), *The psychology of abilities, competencies, and expertise* (pp. 70–92). Cambridge University Press. https://doi.org/10.1017/CBO9780511615801.005

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*, 1-52 https://doi.org/10.1017/S0140525X1500031X

Colom, R., Rebollo, I., Palacios, A., Juan-Espinosa, M., & Kyllonen, P. C. (2004). Working memory is (almost) perfectly predicted by g. *Intelligence*, *32*(3), 277–296. https://doi.org/10.1016/j.intell.2003.12.002

Conway, A. R. A., Macnamara, B. N., & Engle De Abreu, P. M. J. (2013). Working memory and intelligence: An overview. In T. P. Alloway & R. G. Alloway (Eds.), *Working memory: The connected intelligence*. Psychology Press., 13-35

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science*, *19*(1), 51–57. https://doi.org/10.1177/0963721409359277

Diaz Abrahan, V., Shifres, F., & Justel, N. (2019). Cognitive benefits from a musical activity in older adults. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.00652

Dickens, W. T. (2007). What is g? *Brookings*. https://www.brookings.edu/research/what-is-g/

Dickens, W. T., & Flynn, J. R. (2001). Heritability estimates versus large environmental effects: The IQ paradox resolved. *Psychological Review*, *108*(2), 346–369. https://doi.org/10.1037/0033-295X.108.2.346

Elwert, F. (2013). Graphical causal models. In S. L. Morgan (Ed.), *Handbook of causal analysis for social research* (pp. 245–273). Springer Netherlands. https://www.springer.com/gp/book/9789400760936

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates Publishers.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*(2), 211–245. https://doi.org/10.1037/0033-295x.102.2.211

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363–406. https://doi.org/10.1037/0033-295X.100.3.363

Ericsson, K. A., & Moxley, J. H. (2013). Working memory that mediates experts' performance: Why it is qualitatively different from traditional working memory. In *Working memory: The connected intelligence* (pp. 109–136). Psychology Press.

Fuster, J. M. (1997). *The prefrontal cortex: Anatomy, physiology, and neuropsychology of the frontal lobe*. Lippincott-Raven.

Gelding, R. W., Harrison, P. M. C., Silas, S., Johnson, B. W., Thompson, W. F., & Müllensiefen, D. (2020). An efficient and adaptive test of auditory mental imagery. *Psychological Research, 85,* 1201–1220 https://doi.org/10.1007/s00426-020-01322-3

George, E. M., & Coch, D. (2011). Music training and working memory: An ERP study. *Neuropsychologia*, *49*(5), 1083–1094. https://doi.org/10.1016/j.neuropsychologia.2011.02.001

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*(4), 323–343. https://doi.org/10.1037/1082-989X.11.4.323

Hambrick, D. Z., Burgoyne, A. P., & Oswald, F. L. (2018). Domain-general models of expertise: The role of cognitive ability. In P. Ward, J. M. Schraagen, J. Gore, & E. M. Roth (Eds.), *The Oxford handbook of expertise* (pp. 55–84). https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780198795872.001.0001/oxfordhb-9780198795872-e-3

Harrison, P. M. C. (2018). *pmcharrison/psychTestRCAT: PsychTestRCAT v1.0.2*. Zenodo. https://doi.org/10.5281/zenodo.1415358

Harrison, P. M. C. (2020). psychTestR: An R package for designing and conducting behavioural psychological experiments. *Journal of Open Source Software*, *5*(49), 2088. https://doi.org/10.21105/joss.02088

Harrison, P. M. C., Collins, T., & Müllensiefen, D. (2017). Applying modern psychometric techniques to melodic discrimination testing: Item response theory, computerised adaptive testing, and automatic item generation. *Scientific Reports*, *7*(1), 3618. https://doi.org/10.1038/s41598-017-03586-z

Harrison, P. M. C., & Müllensiefen, D. (2018). Development and validation of the Computerised Adaptive Beat Alignment Test (CA-BAT). *Scientific Reports*, *8*(1), 12395. https://doi.org/10.1038/s41598-018-30318-8

Hayduk, L., Cummings, G., Stratkotter, R., Nimmo, M., Grygoryev, K., Dosman, D., et al. (2003). Pearl's d-separation: One more step into causal thinking. *Structural Equation Modeling: A Multidisciplinary Journal*, *10*(2), 289–311. https://doi.org/10.1207/S15328007SEM1002_8

Hayduk, L. A., & Littvay, L. (2012). Should researchers use single indicators, best indicators, or multiple indicators in structural equation models? *BMC Medical Research Methodology*, *12*(1), 159. https://doi.org/10.1186/1471-2288-12-159

Hille, A., & Schupp, J. (2015). How learning a musical instrument affects the development of skills. *Economics of Education Review*, *44*, 56–82. https://doi.org/10.1016/j.econedurev.2014.10.007

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

Hommel, B., & Colzato, L. S. (2017). The grand challenge: Integrating nomothetic and ideographic approaches to human cognition. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00100

Hyde, K. L., Lerch, J., Norton, A., Forgeard, M., Winner, E., Evans, A. C., & Schlaug, G. (2009). The effects of musical training on structural brain development: A longitudinal study. *Annals of the New York Academy of Sciences*, *1169*, 182–186. https://doi.org/10.1111/j.1749-6632.2009.04852.x

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences: An introduction*. Cambridge University Press. https://doi.org/10.1017/CBO9781139025751

James, C. E., Zuber, S., Dupuis-Lozeron, E., Abdili, L., Gervaise, D., & Kliegel, M. (2020). Formal string instrument training in a class setting enhances cognitive and sensorimotor development of primary school children. *Frontiers in Neuroscience*, *14*. https://doi.org/10.3389/fnins.2020.00567

Jones, D. S., & Podolsky, S. H. (2015). The history and fate of the gold standard. *Lancet (London, England)*, *385*(9977), 1502–1503. https://doi.org/10.1016/S0140-6736(15)60742-5

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., & Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, *47*(1), 1–26. https://doi.org/10.18637/jss.v047.i11

Kline, R. (2012). Model fit and model selection in structural equation modelling. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling*. Guilford Press., 209-231

Lehmann, A. C., Sloboda, J. A., Woody, R. H., & Woody, R. H. (2007). *Psychology for musicians: Understanding and acquiring the skills*. Oxford University Press.

Lewis, J. W., & Escobar, L. A. (1986). Suppression and enhancement in bivariate regression. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *35*(1), 17–26. https://doi.org/10.2307/2988294

Li, Y., & Geary, D. C. (2017). Children's visuospatial memory predicts mathematics achievement through early adolescence. *PLOS ONE*, *12*(2), e0172046. https://doi.org/10.1371/journal.pone.0172046

Lorenzo, O., Herrera, L., Hernández-Candelas, M., & Badea, M. (2014). Influence of music training on language development. A longitudinal study. *Procedia - Social and Behavioral Sciences*, *128*, 527–530. https://doi.org/10.1016/j.sbspro.2014.03.200

McNally, R. J. (2016). Can network analysis transform psychopathology? *Behaviour Research and Therapy*, *86*, 95–104. https://doi.org/10.1016/j.brat.2016.06.006

Meinz, E. J., & Hambrick, D. Z. (2010). Deliberate practice is necessary but not sufficient to explain individual differences in piano sight-reading skill: The role of working memory capacity. *Psychological Science*, *21*(7), 914–919. https://doi.org/10.1177/0956797610373933

Moffa, G., Catone, G., Kuipers, J., Kuipers, E., Freeman, D., Marwaha, S., et al. (2017). Using directed acyclic graphs in epidemiological research in psychosis: An analysis of the role of bullying in psychosis. *Schizophrenia Bulletin*, *43*(6), 1273–1279. https://doi.org/10.1093/schbul/sbx013

Morgan, S. L. (2013). *Handbook of causal analysis for social research*. Springer Science & Business Media.

Mosing, M. A., Madison, G., Pedersen, N. L., Kuja-Halkola, R., & Ullén, F. (2014). Practice does not make perfect: No causal effect of music practice on music ability. *Psychological Science*, *25*(9), 1795–1803. https://doi.org/10.1177/0956797614541990

Mosing, M. A., Verweij, K. J. H., Madison, G., Pedersen, N. L., Zietsch, B. P., & Ullén, F. (2015). Did sexual selection shape human music? Testing predictions from the sexual selection hypothesis of music evolution using a large genetically informative sample of over 10,000 twins. *Evolution and Human Behavior*, *36*(5), 359–366. https://doi.org/10.1016/j.evolhumbehav.2015.02.004

Mrazik, M., & Dombrowski, S. C. (2010). The neurobiological foundations of giftedness. *Roeper Review: A Journal on Gifted Education*, *32*(4), 224–234. https://doi.org/10.1080/02783193.2010.508154

MacGregor, C., Andrade, P. E., Forth, J., & Frieler, K., Müllensiefen, D. (2022). The Rhythm Ability Test (RAT): A new test of rhythm memory in children and adults. (Manuscript in preparation).

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

Münte, T. F., Altenmüller, E., & Jäncke, L. (2002). The musician's brain as a model of neuroplasticity. *Nature Reviews. Neuroscience*, *3*(6), 473–478. https://doi.org/10.1038/nrn843

Okada, B. M., & Slevc, L. R. (2018). Individual differences in musical training and executive functions: A latent variable approach. *Memory and Cognition*, *46*(7), 1076–1092. https://doi.org/10.3758/s13421-018-0822-8

Overmann, K., Wynn, T., & Coolidge, F. (2013). The evolution of working memory. In T. P. Alloway & R. G. Alloway (Eds.), *Working memory: The connected intelligence*, 37-60. Psychology Press.

Patel, A. (2011). Why would musical training benefit the neural encoding of speech? The OPERA hypothesis. *Frontiers in Psychology*, *2*, 142. https://doi.org/10.3389/fpsyg.2011.00142

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.

Pearl, J. (2009). *Causality*. Cambridge University Press.

Pearl, J. (2017). *The eight pillars of causal wisdom*. UCLA: West Coast Experiments, Los Angeles.

Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal inference in statistics: A primer*. John Wiley & Sons.

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Allen Lane.

Pham, A. V., & Hasson, R. M. (2014). Verbal and visuospatial working memory as predictors of children's reading ability. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *29*(5), 467–477. https://doi.org/10.1093/arclin/acu024

Plomin, R. (2018). *Blueprint: How DNA makes us who we are*. Penguin United Kingdom.

Plomin, R., DeFries, J. C., Knopik, V. S., & Neiderhiser, J. M. (2016). Top 10 replicated findings from behavioral genetics. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *11*(1), 3–23. https://doi.org/10.1177/1745691615617439

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://www.jstatsoft.org/v48/i02/

Ruthsatz, J., Detterman, D., Griscom, W. S., & Cirullo, B. A. (2008). Becoming an expert in the musical domain: It takes more than just practice. *Intelligence*, *36*(4), 330–338. https://doi.org/10.1016/j.intell.2007.08.003

Salthouse, T. A. (2009). When does age-related cognitive decline begin? *Neurobiology of Aging*, *30*(4), 507–514. https://doi.org/10.1016/j.neurobiolaging.2008.09.023

Salthouse, T. A., Babcock, R. L., & Shaw, R. J. (1991). Effects of adult age on structural and operational capacities in working memory. *Psychology and Aging*, *6*(1), 118–127. https://doi.org/10.1037/0882-7974.6.1.118

Schellenberg, E. G. (2004). Music lessons enhance IQ. *Psychological Science*, *15*(8), 511–514. https://doi.org/10.1111/j.0956-7976.2004.00711.x

Schellenberg, E. G. (2020). Correlation = causation? Music training, psychology, and neuroscience. *Psychology of Aesthetics, Creativity, and the Arts*, 475–480. https://doi.org/10.1037/aca0000263

Schulze, K., Zysset, S., Mueller, K., Friederici, A. D., & Koelsch, S. (2011). Neuroarchitecture of verbal and tonal working memory in nonmusicians and musicians. *Human Brain Mapping*, *32*(5), 771–783. https://doi.org/10.1002/hbm.21060

Seither-Preisler, A., Parncutt, R., & Schneider, P. (2014). Size and synchronization of auditory cortex promotes musical, literacy, and attentional skills in children. *Journal of Neuroscience*, *34*(33), 10937–10949. https://doi.org/10.1523/JNEUROSCI.5315-13.2014

Shah, P., & Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing: An individual differences approach. *Journal of Experimental Psychology: General*, *125*(1), 4–27. https://doi.org/10.1037/0096-3445.125.1.4

Shrier, I., & Platt, R. W. (2008). Reducing bias through directed acyclic graphs. *BMC Medical Research Methodology*, *8*, 70. https://doi.org/10.1186/1471-2288-8-70

Silas, S., Tsigeman, E., Frieler, K., Likhanov, M., Gelding, R., Kovas, Y., et al. (2022). The Jack and Jill Adaptive Working Memory Task: Construction, calibration and validation. *PLoS ONE*, *17*(1): e0262200. https://doi.org/10.1371/journal.pone.0262200

Simmering, V. R., & Perone, S. (2013). Working memory capacity as a dynamic process. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00567

Spearman, C. (1904). 'General intelligence,' objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–292. https://doi.org/10.2307/1412107

Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G., & Richardson, T. (2000). *Causation, prediction, and search*. MIT Press.

Suárez, L., Elangovan, S., & Au, A. (2016). Cross-sectional study on the relationship between music training and working memory in adults. *Australian Journal of Psychology*, *68*(1), 38–46. https://doi.org/10.1111/ajpy.12087

Talamini, F., Altoè, G., Carretti, B., & Grassi, M. (2017). Musicians have better memory than nonmusicians: A meta-analysis. *PLOS ONE*, *12*(10), e0186773. https://doi.org/10.1371/journal.pone.0186773

Talamini, F., Carretti, B., & Grassi, M. (2016). The working memory of musicians and nonmusicians. *Music Perception*, *34*(2), 183–191. https://doi.org/10.1525/mp.2016.34.2.183

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

Tarr, B., Launay, J., & Dunbar, R. I. M. (2014). Music and social bonding: "Self-other" merging and neurohormonal mechanisms. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01096

Tian, J., & Pearl, J. (2013). Causal discovery from changes. *ArXiv:1301.2312 [Cs]*. http://arxiv.org/abs/1301.2312

Vergauwe, E., Barrouillet, P., & Camos, V. (2010). Do mental processes share a domain-general resource? *Psychological Science*, *21*(3), 384–390. https://doi.org/10.1177/0956797610361340

Vinkhuyzen, A. A. E., van der Sluis, S., Posthuma, D., & Boomsma, D. I. (2009). The heritability of aptitude and exceptional talent across different domains in adolescents and young adults. *Behavior Genetics*, *39*(4), 380–392. https://doi.org/10.1007/s10519-009-9260-5

Visser, B. A., Ashton, M. C., & Vernon, P. A. (2006). Beyond g: Putting multiple intelligences theory to the test. *Intelligence*, *34*(5), 487–502. https://doi.org/10.1016/j.intell.2006.02.004

Vock, M., & Holling, H. (2008). The measurement of visuo-spatial and verbal-numerical working memory: Development of IRT-based scales. *Intelligence*, *36*(2), 161–182. https://doi.org/10.1016/j.intell.2007.02.004

Williams, K. E., Barrett, M. S., Welch, G. F., Abad, V., & Broughton, M. (2015). Associations between early shared music activities in the home and later child outcomes: Findings from the longitudinal study of Australian children. *Early Childhood Research Quarterly*, *31*, 113–124. https://doi.org/10.1016/j.ecresq.2015.01.004

Williams, L. J. (2012). Equivalent models: Concepts, problems, alternatives. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 247–260). The Guilford Press.

Yang, H., Ma, W., Gong, D., Hu, J., & Yao, D. (2014). A longitudinal study on children's music training experience and academic development. *Scientific Reports*, *4*. https://doi.org/10.1038/srep05854

Zuk, J., Andrade, P. E., Andrade, O. V. C. A., Gardiner, M., & Gaab, N. (2013). Musical, language, and reading abilities in early Portuguese readers. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00288

Appendix

Descriptive Statistics for Participant Age, MT Measure, VSWM Measures, and MWM Measures

Measure	Mean	*SD*	Min	Max	Range	Skewness	Kurtosis
Age	26.44	7.68	18	50	32	.97	.15
MT	27.88	12.61	7	48	41	−.07	−1.33
BDS	−.01	1.13	−2.79	2.73	5.51	−.18	.29
MUF	.15	.91	−2.01	2.93	4.94	−.03	.19
JaJ	.63	.97	−1.50	2.36	3.87	−.36	−.48
RAT	.63	.81	−2.42	1.91	4.33	−1.54	3.33
PIAT	.75	1.50	−2.69	4.00	6.69	.52	−.12
MDT	.22	1.09	−3.15	2.20	5.36	−.45	−.21