ROYAL SOCIETY OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

Research

Cite this article: MacRitchie J, Chmiel A, Stevens CJ, Dean RT. 2025 Progressively learned musical ability predicts cognitive transfer in older adult novices: a 12-month musical instrument training programme. *R. Soc. Open Sci.* **12**: 251022.
https://doi.org/10.1098/rsos.251022

Received: 6 June 2025
Accepted: 20 October 2025

Subject Category:
Psychology and cognitive neuroscience

Subject Areas:
psychology

Keywords:
cognitive training, older adults, musical instrument learning, improvisation

Author for correspondence:
Jennifer MacRitchie
e-mail: j.macritchie@sheffield.ac.uk

Supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.8184491.

THE ROYAL SOCIETY PUBLISHING

Progressively learned musical ability predicts cognitive transfer in older adult novices: a 12-month musical instrument training programme

Jennifer MacRitchie[1,2], Anthony Chmiel[1,3], Catherine J. Stevens[1] and Roger T. Dean[1]

[1]The MARCS Institute for Brain, Behaviour, and Development, Western Sydney University, Penrith, Australia
[2]Department of Music, The University of Sheffield, Sheffield, UK
[3]The University of Sydney, Sydney Conservatorium of Music, Sydney, New South Wales, Australia

JM, 0000-0003-4183-6552; AC, 0000-0003-3294-0534; CJS, 0000-0002-7558-2717; RTD, 0000-0002-8859-8902

Musical instrument training is an often-suggested candidate for cognitive training in older adults. Studies are typically short-term, with little opportunity to explore different trained music-making activities (e.g. improvisation) or how progression in music-specific learning may affect cognitive outcomes. This long-term study (12 months training, six months follow-up) contributes the first comparison of music-making activity and instruments for healthy older adult novices, evaluating how different conditions may affect music learning (measured quantitatively by objective computational means) and how this transfers to domain-general cognitive and motor skills. Sixty-eight participants experienced both types of music-making activity (replication versus improvisation) and instrument (piano keyboard versus iPad-based ThumbJam) through four three-month blocks delivered online. We have previously published our investigation of the participants' music learning, with the biggest improvements in melodic discrimination tasks associated with improvisation training. This article uses our repeated measures design of domain-general cognitive and motor skills to demonstrate that the extent of learning, as evaluated by music-specific perception tests, can predict some

cognitive benefits. Implications are in the design of music teaching and learning tasks for cognitive gain, such that individuals can be supported to develop skills to the best of their ability.

1. Introduction

Musical training as a cognitive intervention across the lifespan has received various attention of late, with correlational evidence to suggest that there is a musician's advantage going into older adulthood, particularly for speech in noise [1], and general cognitive benefits [2]. A positive relationship between sustained arts participation and general well-being in older adulthood [3] as well as increased opportunities for connecting with others in a shared activity [4,5] and the high value of music as a cultural activity make it an attractive potential candidate for cognitive intervention.

The hypothesis that music training leads to domain-general cognitive gain is well-discussed in the literature, with authors pointing towards relatively weak effects of any such far transfer observed both in the specific domain of music training for older adults' executive functions [6] and in general cognitive training literature [7]. Arguments put forward by Schellenberg & Lima [8] suggest that pre-existing factors may determine who takes music lessons and cognitively succeeds, rather than music training itself being the main driver of cognitive change. Musical training is generally 'broad and undefined' when described in intervention studies, and very seldom is the extent of musical learning measured objectively, particularly for older adults. In turn, it becomes difficult to pinpoint the features of any programme that may lead to specific cognitive benefits or provide appropriately selected control programmes other than different types of activities (such as reading or cooking).

Our reported research makes three distinct contributions. First, we expose older adult novices to a long-term music training programme of 12 months duration and follow-up over six months post-training. Previous interventions' weak effects have been suggested to be in part due to length of intervention (three–six months in general in comparison with correlational studies looking at 10+ years of training). An exception is Worschech *et al.*'s and Mack *et al.*'s studies [9,10] which examine the effects of a year-long piano training programme using magnetic resonance imaging (MRI), cognitive (auditory digit span) and motor (Purdue Pegboard) tests at 6 and 12 months. In [9], the authors aimed to examine whether there was a dynamic coupling between cognitive and motor gains; no evidence was found for this. Our study adds to this knowledge by contributing more frequent testing sessions (every three months). In these sessions, we examine measures of executive function that have been identified as good candidates from systematic reviews of the effects of music training programmes for older adults, such as the Trail Making Test [6], a battery of music performance and music perceptual tasks (including the Melodic Discrimination Test, hence MDT [11] and Computerised Adaptive Beat Alignment Test, hence CA-BAT [12]). Our primary focus is on assessing the causal effect between music training conditions and the cognitive outcomes; however, we include psychosocial measures as secondary measures to take into account a holistic view of older adults' health and well-being over the programme.

Second, we have compared the effect of different types of music-performing activities in melodic replication and improvisation (hereafter referred to as tasks) on music learning [13], with the aim of investigating whether aspects of music training design and consequent music learning may influence any cognitive or motor gains. Few music intervention studies examine improvisation, despite there being evidence to suggest that its inclusion may increase music performing ability in children [14] and a correlation between improvisation skills and general music sophistication [15]. A 10-day jazz piano training intervention for older adults teaching improvisational skills based on traditional notation [16] shows some small difference against an inactive control group in task switching as measured by the Stroop test. Improvisation in a three-month intervention for older adults shows enhanced global cognition and spatial working memory, but no significant differences to that of a traditional music learning programme [17]. The crossover design of our reported research programme here is unique within the music training literature and allows the effect of every aspect to be distinguished from others, including distinguishing that of time progression from that of the extent of improvisation or replication learning.

Third, our analyses investigate in depth the relationship between changing musical measures of ability and far transfer measures of domain-general motor skill and cognition. Two short-term interventions [16,18] and a longer-term intervention [10] briefly compare simple music performance

measures (playing a five-note scale in [10]) directly against cognitive outcomes. Until now, this has not yet been given serious attention.

In our programme, dubbed the Active Minds Music Ensemble (AMME), our specific research aims were to assess (i) the effects of cognitive focus on the rate of learning operationalized as differences in trained music-making task (melodic improvisation/replication) and in physical demand via instrument type (piano keyboard/iPad ThumbJam, as detailed below); (ii) the extent to which domain-specific learning outcomes in music may transfer to domain-general cognition; and (iii) the retention of skills six months post-training. Automated techniques were developed specifically for this project to allow a thorough investigation of music learning [19].

Given the positive results of music learning (our first research aim) by our participants [13], we anticipated that a measure of music learning (individualized to participant and session) would be the strongest predictor of any observed domain-general cognitive and motor changes (aims ii and iii) in comparison with the count of blocks of training alone. Our hypotheses were then as follows: if creating new material enhances learning, then participants demonstrate enhancements in cognitive and motor tasks when instruction involves improvisation compared with replication tasks (H1). Given that we are interested in the contribution of improvisation tasks to creative thinking, we included a domain-general measure to establish the transfer of this particular skill into a domain beyond manipulating the trained musical material. Second, we hypothesized that older adults demonstrate enhancements in cognitive and motor tasks when the instrument involves a low rather than high motor requirement (H2).

In [13], we used a two-level factor (an independent variable) representing the extent of improvisation training undertaken and of replication undertaken (referred to as *ctimprep*, counts of improvisation and replication training blocks experienced by a participant) and, more importantly, demonstrated the contribution to participants' musical learning measured in several aspects. In particular, strong evidence showed the impact of improvisation sessions on the development of aural perception as measured by the MDT (which asks whether a listener can distinguish a slightly changed melody from its original). Thus, in order to test the influence of musical learning *per se* (as opposed to the specific and social experience of the music classes), in the present paper, we mainly use the individual MDT scores obtained from the model of the learning data (i.e. a score for each individual at each test session) as a potentially key predictor of any motor and cognitive changes we observed concomitantly. We show here that the musical learning data predicts the majority of such motor and cognitive changes, while the progression of the training sessions alone (the test session count and the more specific *ctimprep*) was unimportant in comparison. In this current article, the primary focus is on presenting the full results of the cognitive and motor test battery over 12 months of repeated measures, as well as a six-month follow-up period.

2. Methods

2.1. Design

To operationalize our research aims, independent variables of task (improvisation and replication) and instrument (keyboard and iPad app ThumbJam) were set up in a crossover design in which each participant experienced every combination of task and instrument, as depicted in figure 1, but with varying relations to the passing of time. We broadly assumed that all participants have the same potential for learning to play, reproduce and improvise, with the expectation that starting levels could vary. Using the individual as their own control and monitoring individual progress allowed an in-depth assessment of this issue.

The two tasks (replication and improvisation) were designed to involve playing single-handed melodies in the range C4–G5. Participants were given aural prompts to learn either through repetition of all notes (replication) or repetition of the three–five note prompt then extension by modes of 14 distinct improvising methods (improvisation, detailed in [13]). The choice of instrument interface (keyboard and ThumbJam) allowed comparison on two distinct interfaces where consecutive pitches were arranged horizontally in a single line. The keyboard interface presented a traditional electronic surface where physical keys had to be pressed down for a note to sound (with varying speeds of pressing the key translating to louder/softer volumes and some vertical movement of the hand required for navigation between white/black notes). ThumbJam instead presented a single row of visual keys on a flat touchscreen, requiring only vertical movement of the hand to produce louder/

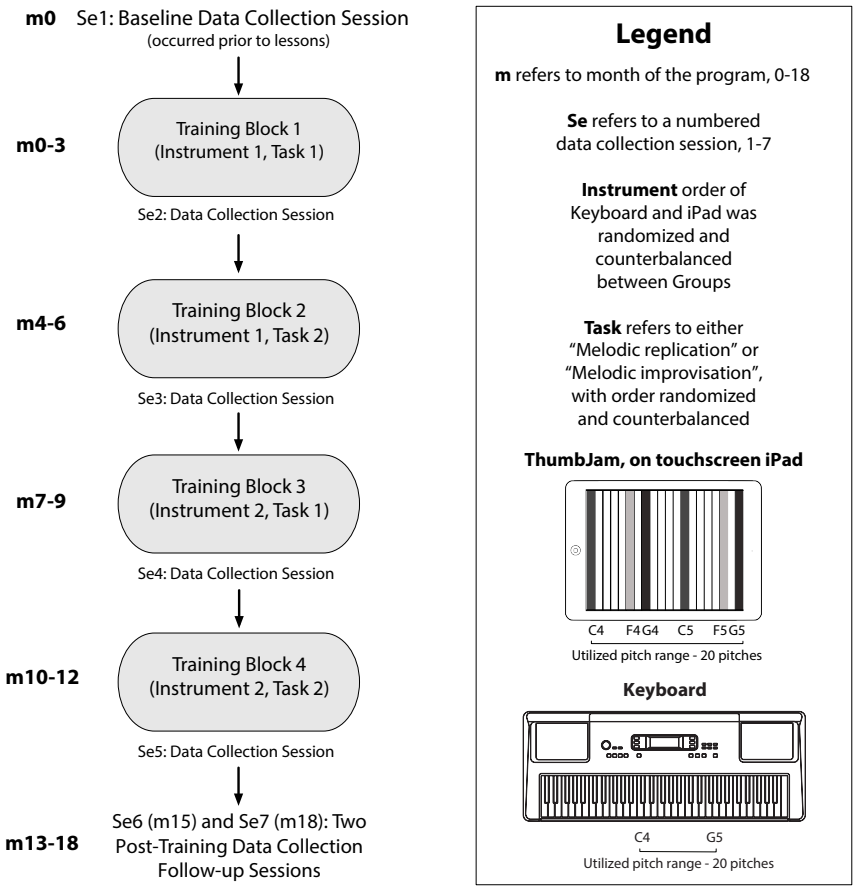

Figure 1. Overview of the AMME experimental crossover design where task and instrument allocation are counterbalanced across groups, and every session has characteristic features.

softer volumes.[1] So as not to present additional cognitive challenges in switching between instruments at each three-month period, the procedure for any one participant was restricted such that they would begin on one instrument for six months, then continue with the other.

This led to four possible training orders for participants (which were distributed as equally as possible across the participant training groups) with one example being: keyboard replication, keyboard improvisation, iPad replication and iPad improvisation. The progression of time was thus distinguishable from the progression of different kinds of study: each block had a combination of those factors different from all others in the sequence.

2.2. Participants

Sixty-eight cognitively healthy participants (aged 65–79, M = 70.3, s.d. = 3.8; 60 female, 8 male) were recruited via social media and adverts through local organizations including the Women's Own Network. Criteria for inclusion were (i) normal to corrected vision and normal to corrected hearing, (ii) no extreme physical ailments in upper extremities such as severe arthritis, (iii) no cognitive impairments as assessed by the shortened version of Addenbrooke's Cognitive Examination (miniACE-III), and (iv) minimal previous musical experience (less than 2 years formal training or playing experience on a musical instrument, excluding voice[2]). The vast majority of participants (71%) reported having no prior musical training experience on an instrument (see the electronic supplementary material for further details). As participants were recruited on a rolling basis over a 2-year period, there was limited ability to match groups *a priori* on characteristics of gender, age, education level or occupational status; however, our data analysis procedure was designed to account for potential individual differences (see below). The characteristics of the participant sample are outlined in table 1.

To clarify the musical sophistication of our participants, Goldsmiths Musical Sophistication Index (Gold-MSI) self-report and tested levels were compared with normative data from the sample of 147 633 participants examined in [21]. As detailed in [13, electronic supplementary

Table 1. Participant demographic information.

demographic or screening measure	mean/s.d./range
age (years)	70.35, 3.76, 65–79
musical training on an instrument, excluding voice (years of formal lessons)	0.36, 0.60, 0–2
cognitive level (mini-ACE)	28.98, 1.06, 26–30
gender	
female	60
male	8
nationality	
Australian	59
Australian and European	6
Filipino	1
Anglo-Indian	1
British	1
education level	
did not complete any school qualification	1(1.5%)
completed first school qualification at 16 years	8(11.8%)
completed second qualification (high school grad)	8(11.8%)
undergraduate degree or professional qualification	27(39.7%)
postgraduate degree	23(33.8%)
still in education	1(1.5%)
occupational status	
in education	1
in part-time employment	8
in full-time employment	3
self-employed	1
retired	55

material], a series of one-sample *t*-tests demonstrated ($p < 0.001$) that our participants were indeed musical novices in all Gold-MSI subscales, barring the 'Emotions' subscale. The Gold-MSI measures were thus our detailed and specific measures of musical learning levels at successive stages of the project, including the outset. The simple descriptor 'years of musical experience' was not used in models, being uninformative and simply an exclusion criterion.

Participant attrition was a total of 13 participants (19.1%) leaving at some point during the 12-month training period. Further details are included in [13, electronic supplementary material].

2.3. Procedure

Our music training programme consisted of hour-long training sessions held once every two weeks. This is described in detail in [13, electronic supplementary materials], including sample files. The programme was co-designed with an experienced piano teacher with skills in improvisation, and all lessons were provided by this same teacher to all groups. Daily practice of approximately 30 min/day was recommended. Adherence to this practice requirement was monitored via interviews discussed below.[3] As such, the 'dosage' of our training programme was considered as 4 h per week (including the lesson plus practice time).[4] Participants were provided with both instruments on loan (one piano keyboard Yamaha PSRE 353 and one Apple iPad running the app ThumbJam) for the duration of their involvement in the programme (the full 18 months which included six months with follow-up assessments). We used both instruments in the data collection procedures online.

The battery of assessment measures was administered online in procedures described below. In most cases, this consisted of the assessor and participant on a *Zoom* call, and the assessor would

guide each individual through the tests, whether they were to be conducted through online webpages, on individuals' own computers, or the iPad or piano keyboards they had as part of the training programme (details of the programs used for data collection are included in the description of each test).

2.4. Randomization and blinding

Our battery of assessment measures is below and was administered by the same blinded assessor to every participant at every time point. Screening measures were only measured at baseline, and the Digit Span Test was only measured at baseline and 12 months, in an effort to keep the battery to a manageable length for participants. The remainder of the tests was administered every three months. The self-report questionnaires were sent via email link to participants in the week running up to their online appointment with author A.C. The test battery order was then quasi-randomized in the following way over four blocks: domain-general tests were presented first, with the exception of the trail Making Test, which was presented third. Blocks two and four were then music-specific tests consisting either of the two tests of music perception (see MDT and CA-BAT below), or the finger tapping and pitch direction test. The order of blocks was randomized for every participant and every session. The order of tests within each block was also randomized.

2.5. Screening measures

Two measures were used to screen for our inclusion/exclusion criteria around cognitive health. A further bespoke measure of music discrimination (pitch direction test, unvalidated) was used primarily to alert the teacher of any additional training needs.[5]

2.5.1. Addenbrooke's Cognitive Examination—III short version (mini-ACE)

The mini-ACE assesses cognitive performance in terms of attention, memory, verbal fluency, language and visuo-spatial abilities [24], providing high diagnostic accuracy for screening of Alzheimer's disease [25] and other dementias [24] with high reliability and consistency over time (Cronbach's α = 0.85, ICC = 0.83 among older adults in long-term care [26], Cronbach's α = 0.83 in dementia [24]). This test was administered via *Zoom*. For the test item requiring the participant to draw on a piece of paper, participants completed the task then held up their drawing to show the assessor. Scores are summed across various tasks to give a total out of 30, with higher scores reflecting better cognitive performance. Two cut-offs are identified in [24] for the use of screening participants for research, the upper being a score of 25, and the lower being a score of 21. A specificity of 1 is obtained at this lower cut-off. All participants in the current study achieved a score of 25 or above to satisfy our inclusion criteria.[6]

2.5.2. Subjective cognitive decline (SCD)

As part of a baseline questionnaire administered through online survey platform *Qualtrics*, participants were asked: 'In your everyday life, how frequently do you have memory complaints/trouble with your memory'. Participants could choose from answer categories 'not at all, rarely, sometimes, frequently or very frequently' to assess any subjective measure of cognitive decline. An open-ended question was provided if participants wanted to clarify this further. Only two participants rated themselves as 'frequently' or 'very frequently'.[7]

2.6. Domain-general tests (repeated measures)

Our battery of cognitive and motor tests included the following measures.[8] Repeated measures for these tests were all at baseline (before exposure to training) and then each subsequent three-month period between blocks of training. Two points at follow-up were recorded (three and six months post-training). Our test sessions were referred to as sessions 1–7. Some tests did not occur in every session. For example, for pragmatic reasons to reduce participant testing time, the Digit Span Test (forward only) was only tested at baseline and 12 months (test sessions 1 and 5).

2.6.1. Digit Span Test—forwards only (0 and 12 months, tested over *Zoom*)

The Digit Span Test from the Wechsler Adult Intelligence Scale IV [29] is a test of working memory regularly used in music intervention studies [6] with a retest reliability of 0.83 and internal consistency reliability of 0.93 [30]. This test was administered via a pre-recorded audio file where the digits in progressively longer strings were spoken aloud with 1 s intervals between digits. Scoring of the test is the number of digits in the string the participant correctly remembers. Two incorrect responses at one string length result in termination of the test. Different sets of numbers were used at the baseline and 12-month mark. The decision to only include the forward version of this test, and at a reduced number of time points in the data collection programme, was due to the low effect size and non-significant result of Rogers & Metzler-Baddeley's meta-analysis [6], as well as the narrow effects observed in wider working memory literature [7] and a need to decrease the length of the test battery for participants. This behavioural test of working memory may reflect an increase in cognitive efficiency rather than cognitive capacity (see [7] for further discussion).

2.6.2. Trail Making Test (0, 3, 6, 9, 12 months, tested on participant iPad)

The Trail Making Test (TMT) assesses participants' visuo-motor skills, sequencing, processing speed and cognitive flexibility [31] and has previously been used to demonstrate cognitive gains as a result of short-term music-training interventions with older adults [4,32] and as a consistent measure of visuo-spatial attention and cognitive switching in music interventions [6]. This task was administered via the NeuRa iPad app (https://neura.edu.au/resources-tools/apps/trail-making-test), which has been validated against pen and paper data collection.[9] Using the app to collect data automates the timing and collection of results. To guard against missing data from technical failure, the assessor requested the participant show the time on screen to them via *Zoom* as soon as each part had been completed. Use of data from this manual data-logging approach was required in 0% of tasks. Three participants were excluded from this dataset as they had been undertaking the test themselves independent of the test sessions (evident by data logs within the *NeuRA* app).

In the TMT, part A consists of an onscreen display of 25 numbered circles whereby participants tap each number consecutively to join them up in a line as fast as possible. Part B is more challenging in terms of executive control and visual search: it consists of 25 circles including both numbers and letters. Participants have to tap the circles onscreen in ascending order, alternating between the numbers and letters, again to be completed in the shortest time possible. The TMT part A has been validated primarily for visuo-perceptual abilities and part B for working memory and cognitive flexibility (referred to as either task-set switching or attentional-set shifting) [33], with high reliability over conditions (parts A and B, both pen and paper and digital delivery ICCs p = 0.90–0.95) [34], although it is susceptible to practice effects in short time periods [31]. Scores are the time taken to complete parts A and B, with lower scores reflecting better performance. Additionally, the difference score delta (time for part B – time for part A) attempts to isolate the part of the score attributable to performance on cognitive flexibility.

2.6.3. Alternate Uses Task (0, 3, 6, 9, 12 months, tested over *Zoom*)

The Alternate Uses Task (AUT: [35]) was chosen to assess domain-general skills in generating ideas; we hypothesized this could be affected by exposure to training in musical improvisation, i.e. generating new musical ideas. The AUT asks participants to name as many uses as possible for a household object that are not its original intended use, within a 2 min period. For instance, if the given object is a shoe and its original intended use is as footwear, alternative use examples could be to use the shoe as a flowerpot or as a drinks receptacle (the 'shoey', representing this, was a popular answer among our Australian participants, being a long-standing piece of local slang). This test was administered via *Zoom*. During testing, the assessor provided participants with the object's original intended use and reminded participants that highly similar alternative uses would only be counted as one use. The object given at each time point was randomized for all participants, and participants were given a different object at every time point.

To ensure an instructional-scoring fit (to maximize reliability [36,37]), as our instructions to participants had been to name as many unorthodox uses for an object as possible rather than as many unusual uses for an object, the sum of such generated uses was taken as the test measure. This represents a participant's fluency in generating alternative uses. This measure was considered

most appropriate for our uses, as it would be the most analogous to generating new musical material. A sample of the data was screened by authors J.M. and A.C. to generate a data cleaning approach. Subsequently, another research assistant removed duplicate uses, vague phrases (e.g. simply stating 'to be used in a creative way') or uses that would not be physically possible (e.g. to fly to the moon on a pencil). This data cleaning was again checked by authors J.M. and A.C. to prevent an overzealous approach in culling unusual alternative uses.

2.7. Self-report questionnaires

All self-report questionnaires were administered as one batch at each testing session via a *Qualtrics* survey.

2.7.1. Shortened Disabilities of Arm, Shoulder and Hand, short version (QuickDASH)

The QuickDASH [38] is an 11-item self-report questionnaire based on the original 30-item DASH that asks participants to rate over the past week the severity of symptoms in their shoulder, arm and hand (e.g. weakness, tingling or pain) and ability to perform activities of daily living (e.g. preparing a meal, carrying a shopping bag or briefcase or pushing open a heavy door). As playing a musical instrument is a motor task, the QuickDASH gives a baseline and pre-/post-score measurement of participants' general physical function. Each item has five possible responses, with a single scale-based score calculated from the individual items (possible scores 0–100, with higher scores reflecting a greater degree of disability). The minimal detectable score and minimal clinically relevant score are 11. The full-length DASH has high internal consistency (reported Cronbach's α = 0.96) with high reliability and validity for people with disorders of upper extremities [39]. The QuickDASH offers similar discriminant ability, cross-sectional and test–retest reliability as the full DASH [38].

2.7.2. General self-efficacy

The general self-efficacy (GSE) scale is a 23-item self-report questionnaire designed to assess an individual's agency in terms of the degree of belief that one's actions are responsible for a successful outcome [40]. Questions include rating agreement with 'When I make plans, I am certain I can make them work' with a 14-point Likert scale. An aggregate score of the individual items is used to assess the degree of self-efficacy (with higher scores reflecting a greater degree of self-efficacy). In order to minimize the different resolution of rating scales across the questionnaires, we adapted this scale to use a 7-point Likert scale rather than the original 14 points. The GSE has two sub-scales: general self-efficacy and social self-efficacy (reported Cronbach's α = 0.86 and 0.71, respectively [40]) to distinguish between beliefs on entering new situations from the ability to foster relationships with others.

2.7.3. Basic Psychological Needs Scale

The Basic Psychological Needs Scale (BPNS) is a 21-item self-report questionnaire examining the psychological needs and satisfaction of individuals in domain-general situations primarily around constructs of autonomy, relatedness and competence [41] (with reported Cronbach's α = 0.69, 0.86, 0.71, respectively, and 0.89 for the overall index [42]). Self-determination theory has been used to frame the well-being benefits of music participation [43] with Hallam *et al*. [44] finding a small significant positive difference between older adults who participate in music versus non-music groups on the overall measure and relatedness subscale.

2.7.4. Socio-cultural activity estimations

We asked participants to estimate a percentage of activity in the waking day spent on eight categories. Participants were asked to rate these online with a slider moving from 0 to 100 per cent. All eight activities did not have to add up to 100% of the waking day. Categories rated were talking with others; engaging with music; physical activities; cultural activities outside the home; reading books; using handheld devices (e.g. smartphone or tablet) for non-reading activities; watching TV or films; other mentally stimulating activities.

2.8. Music-specific tests

2.8.1. Goldsmith's Musical Sophistication Index (0, 3, 6, 9, 12, 15, 18 months, assessed via online web browser)

The MDT [11] and the CA-BAT [12] from the Gold-MSI battery were used to determine aspects of music perception skills whereby the MDT would relate to pitch perception and the CA-BAT to beat perception. The MDT has reported strong construct validity and good test–retest reliability of $r = 0.79$ for 10 items, rising to $r = 0.88$ for 20 items (simulating Pearson correlation between ability estimates and true ability scores [11]). The CA-BAT has lower but acceptable levels of test–retest reliability ($r = 0.51$ after 15 items and $r = 0.67$ for 25 items) [12]. These tests were used as a measure of musical learning attained through the instrument training programme, as the tasks involved replicated trained tasks in perceiving pitches, durations and sequences of notes. The MDT measures whether a participant can distinguish two similar melodies presented in rapid succession. The CA-BAT asks participants to distinguish whether a metronome pulse is aligned (or not) with the beat of the accompanying music extract.

These tests were administered via an online server running R, with the tests themselves running in PsychTestR by Goldsmith's researchers [45]; author A.C. created instructions to use PsychTestR to create online studies, available at [46]. Both of these computerized tests take an adaptive approach whereby the item selection is adapted to the participant's previous responses. An ability estimate is calculated from the item response model on participant performance. This score ranges from −4 to +4. Listening environment (volume, use of headphones/speakers) and participant attention to the study in an online delivery format was monitored by the assessor via the *Zoom* call.

2.8.2. Finger tapping test (0, 3, 6, 9, 12, 15, 18 months, assessed via piano keyboard and *Max* patch)

We used a modification and expansion of the finger tapping test to measure general manual dexterity [47,48] (used in studies of music interventions in older adults [49]), expanding to test (i) single taps with the right index finger on a single white key (middle C), and (ii) alternating taps with the right hand index and middle fingers on adjacent white keys (C and D). Each task required the participant to produce the maximum number of taps possible in 10 s duration. Considering motor activities essential for keyboard melody replication and improvisation, we anticipated the alternating taps to be more of a trained motor task than the repetition of a single note. These tests were conducted on the piano keyboard and administered online by the use of a bespoke *Max*/*MSP* patch using a 3 s countdown to a starting sound. The patch then counted the number of taps made on the piano keys in question in the 10s duration. At each testing session, participants performed the single finger task three times in succession (with breaks allowed between attempts), followed by the dual finger task three times in succession, again with breaks in between allowed. For each of the single and dual finger tasks, the attempt with the highest recorded number of taps was used.

To account for any technical issues in recording the data on participants' own computers, a video recording over *Zoom* was also made of the task, and if required the assessor made a manual count of taps by viewing a 30% speed version of the video. Less than 5% of attempts required the manual counting approach.[10] A series of comparisons was performed between data collected by *Max*/*MSP* and via the manual counting approach. In each case, the manual counting approach produced the same count as the *Max*/*MSP* patch, supporting the accuracy of this complementary method. It should be noted that our version of this test is currently unvalidated as it is modified from the original, with the modifications in trial number being made primarily to counter participant fatigue, and the difference in tapping instrument (the piano keyboard) made to be more closely related to the trained instrument in the intervention.

2.8.3. Musical performance tests (0, 3, 6, 9, 12 months, recorded directly onto instrument via MIDI protocol)

Musical performance tests were recorded at every three-month period. Due to the volume of data collected and analyses necessary to fully distinguish the music-specific learning effects, these have been reported extensively and analysed in [13].

2.8.4. Participant interviews (0, 3, 6, 9, 12 months, conducted over *Zoom*)

Participants were interviewed by author J.M. or R.T.D. on their experiences of the lessons. A subset of this data has been reported in [50], with analyses still in progress on the remainder. Semi-structured

interviews served a dual purpose of sustaining participant engagement with the training programme, identifying any training needs, while also collecting relevant data on each individual's motivation, adherence to practice stipulations, their expectations versus reality of the training programme and their attitudes towards and competence with technology.

2.9. Data analysis

A key predictor we wished to assess in models of motor and cognitive change went beyond the independent variable that represents the extent of improvisation and replication training, the variable termed *ctimprep*: i.e. we wanted instead to use an objective measure of musical learning as a predictor. For this purpose, we used the MDT values (as described above) at each assessment point, by individual.

Nevertheless, *ctimprep* remained useful for assessing possible distinctive roles of the improvisation and replication training in observed cognitive, motor and other changes. In *ctimprep* (*a–b*), *a* is the count of improvisation blocks of training undertaken and *b* that of replication, at a particular step in the teaching, each ranging from 0 to 2 (for more detail on the *ctimprep* predictor see [13]). Since *ctimprep* is unchanging across test sessions 5–7, in some cases, we separated the analyses of test sessions 1–5 from those of 5–7; we also considered when appropriate an additional variable, the total count of test sessions experienced (and hence the progression of time), ranging from 1 to 7, 1–5 representing the test sessions closely surrounding the four blocks of teaching, 6–7 the two subsequent three-month blocks after teaching. Using this predictor as a factor variable (rather than numeric) allows for any expression of progression or regression of learning due to the elapse of time by three-month training block to be shown by the relevant coefficient. Note that, as shown in figure 1, test sessions 1–7 are represented as m0–m3 . . . m18 (showing how many months of involvement in the experiment have passed, and where teaching was m0–m12).

As we show, the use of the music learning measure (MDT) as a predictor allows us to provide direct evidence whether the music training could act on motor and cognition functions through any consequent music learning. Note that for some of our tests, notably improvising and replicating ability, we chose not to attempt a 'zero-time' baseline value before the first training block (and so there are for these tests no data for *ctimprep* = 0–0). During the training blocks, i.e. across test sessions 1–5, the session count is not only an index of time *per se*, but also of time of social exposure in the *Zoom* sessions to the other participants and the teacher and researchers; furthermore, it encompasses the progressive incrementing of the number of times any individual cognitive or motor assessment had been undertaken, hence if there were significant improvements in performance due to this repetition, they would appear as an effect of the session parameter. (As detailed in the preceding work on musical learning, and again observed here and described below, there were minimal such effects, hence none of these correlates of time progression were significant influences on the cognitive and motor changes.)

We model each individual test dataset in the same Bayesian way (see the next section): after assessing the possible progressive changes with session and in relation to the cumulation of improvisation and replication study using a simple descriptive model *Score ~ Session* + (*Session* + *Group* | *pid*) (not shown), we then consider possible mediation of the effects by musical learning *per se* (mechanistic models, corresponding figures part b). We use the data (by individual by session) from the MDT, done in formal tests during every test session 1–7, as our representation of each individual's extent of musical learning, since this represents aural cognitive skill.

On consideration of the musical training assessments we made, it is clear that there are none that specifically and solely test motor skills, although participants identified that they progressed in this respect on both instruments (see [13]). Rather than model the motor tests undertaken in the formal test sessions on the basis of such subjective 'fluency' on the iPad and keyboard, we therefore also assessed the objective MDT data again as a predictor of the motor developments. Our main analytical model form in the statistical platform R, used for each graph shown when data were available from all seven test sessions, was: *Motor/CognitiveOutcome* ~ *MDTscore* + *Session* + (*Session* + *Group*|*pid*) where *pid* refers to a unique participant identifier, and *Group* to the participant group (1–10). We added *ctimprep* when we sought to distinguish the roles of the improvisation and replication training. When data were only available from some of the test session time periods (m0-m15), simplified models were necessary, as specified in the Results.

For the self-report questionnaire analyses DASH, BPNS, GSE and activity divisions, simple descriptive models of the form *Score ~ Session* + *(Session* + *Group* | *pid*) were used to assess whether there were any strongly evidenced changes with music training.

As mentioned, the *Group* variable was routinely included in the models to account for any variation across the initial demographic information (including age, education level and previous musical experience). However, the effect of the *Group* variable rarely made enough of an impact to be retained during model selection for optimized models, particularly over and above the *pid* group effect over the whole sample.

2.10. Statistical analysis

In R, we used Bayesian regression modelling in the package brms, with dissection of the effect sizes within complex models by means of the marginal effects package. We have provided an in-depth discussion of the benefits of this approach in [13, electronic supplementary material], and the interested reader is referred there. Model selection was primarily based on the consideration of the observed error of the modelled values and the Bayesian R^2 between model and data.

In brief, the Bayesian approach is highly conservative, providing unambiguous null and positive hypothesis testing using 'evidence ratios' (hereafter e.r.). The evidence ratio is the ratio between the proportion of the modelled probability distribution of a predictor that is in favour of the specific hypothesis about that predictor, and that which is against. So a 95% : 5% distribution between 'in favour' and 'contrary' components provides an evidence ratio of 19 (and is partially analogous to a frequentist $p = 0.05$). For a unidirectional hypothesis, this is termed a 'strong' evidence ratio, while for a bidirectional hypothesis, a ratio of 39 is required for that designation [51]. Thus, our usage of 'strong evidence' is a generally accepted classification rather than an interpretation on our part. Note that the evidence ratio is quite different from a Bayes factor, which is normally an attempt to assess the degree to which one model of a dataset is to be preferred to another, commonly used within frequentist analysis, and not used here.

The Bayesian approach provides an in-depth assessment of the data (by extensive Markov chain Monte Carlo (MCMC) sampling) resulting in distributional descriptions of all the modelled and predictor variables with clear 'Bayesian credibility' intervals attached to the distributions. These are the detailed estimates of the range of plausible values of the parameters, rather than point estimates and confidence intervals (whose interpretation is much more complex). Bayesian analysis also provides a 'posterior' statistical model of the studied data that can be used to predict the data both at precisely the conditions used and also at any combination of conditions that is within the ranges used (so-called 'counterfactual' predictions). In [13], we illustrate counterfactual predictions that support the directly deduced evidence of a lack of effect of the progression of time in the group social conditions upon the music learning. In the present analyses, again, not only do models show minimal coefficients on counts of time units *per se* (as opposed to counts of improvisation or replication training) but also counterfactuals in which replication and improvisation block counts are set at zero while time counts progress predict no convincing cognitive or motor changes. The general implication that it is music learning that is the dominant influence on other changes in our participants holds whenever tested in the present dataset. Some additional comments on the Bayesian modelling are provided in the present electronic supplementary material.

3. Results

A detailed assessment of music-specific tests has been reported in our previous publications [13]. In summary of this, Bayesian modelling showed that melodic discrimination (MDT) and music performance measures assessing abilities of replication and improvisation were learned and seemingly progressively, while beat detection (CA-BAT) and rhythmic precision in replication music performance measures were not. These skills were retained over a six-month follow-up period. Improvisation teaching was the bigger predictor of melodic discrimination, and replication teaching of replication performance. Thus, we used the individual and session-specific values of MDT as predictors in our models here.

3.1. Finger tapping tests

There was no change in maximum single finger tapping rate across the testing sessions, while dual tapping was clearly enhanced. A good model of dual tapping was obtained (electronic supplementary

material, model S1: model error sigma 7.85, which is reasonable considering a mean maximum dual tapping value of 66; and Bayesian R^2 of 0.80), with strong evidence for positive coefficients on MDT score (evidence ratio 28.0), and sessions 5–7 versus session 1 (e.r. 19.8, 46.1, 88.6, respectively). Figure 2 shows the results.

3.2. Domain-general tests

3.2.1. Digit Span Test

In posterior predictions by Session (m0 versus m12, as only two measurements of this parameter were made), there was evidence for improvement by end of training (m12), but with a weak reported evidence ratio (8.1). However, there was very strong evidence for an influence of the individual MDT scores (e.r. 160.3) as shown in figure 3. (The good model obtained is shown as electronic supplementary material, model S2, with sigma 1.08, and Bayesian R^2 0.71.) In this case, the impact of the MDT score was dominantly evidenced, while it was again clear that improvement occurred during training (i.e. between m0 and m12).

3.2.2. Trail Making Test (TMT)

3.2.2.1. Trail Making Test part A

Figure 4 shows the result from our standard modelling approach (electronic supplementary material, model S3, which was again optimal though less powerful than the models presented above). The session counts in this and other graphs represent the effect of the relevant mean conditions and are not simply the consequence of the progression of time and social exposure. MDT score can explain a good part of the response, and it was also noted that timings were lower after keyboard than iPad blocks (with a weak e.r. of 12.4). Here, the fact that Session 3 was strongly evidenced as quicker in TMT A than the commencement values shows that a learning effect was achieved once there had been one session of each of improvisation and replication training; separate analyses using *ctimprep* in the model confirmed this.

3.2.2.2. Trail Making Test part B

Here the model with *ctimprep* was detectably better than that with Session; because the data involved include all the information in Session, it can still be used to provide posterior predictions of the dependencies on Session (or *ctimprep*) of the outcome variable. The model is shown in electronic supplementary material, model S4, and figure 5 shows the key results. While the positive effect of MDT score as predictor of TMT B timing was visible, it was poorly evidenced (e.r. 3.14), and that of keyboard versus iPad even more weakly (e.r. 2.13). Session effects in figure 5 were strong, and there was even strong evidence of enhancement of the timing improvement during sessions 5–7 post-training (potentially representing consolidation and paralleling the weak trend in TMT A, figure 4). We again assessed possible impacts of age and subjective cognitive decline (SCD) in a range of models and found unsystematic effects with low evidence ratios.

It is normal to consider separately improvements (reductions in time) in the difference between TMT parts A and B (delta). This TMT delta outcome showed central values by session from 50 to 42.5, and within these smaller numbers, very similar patterns to those of figure 5, including the lack of effect of age and SCD.

3.2.3. Alternate Uses Task

We chose to model the numerical diversity of envisaged uses, providing at least a restricted index of inventiveness if not creativity at large. We included both item (the particular object named for creative consideration) and participant group (random) effects and assessed the population (fixed) effect of the objects also. Among the 10 objects used, 'brick' appeared as the least capable of triggering use conceptions, and its negative coefficient as a population effect (−0.99) was strongly evidenced (e.r. 56.7). There were no systematic changes across the whole training schedule, though there was strong evidence (e.r. 147.2) that after one block of improvisation training, performance on the task was higher

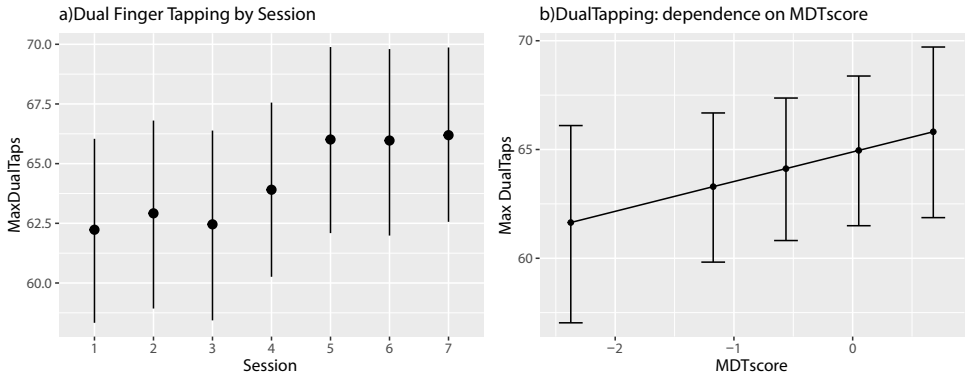

Figure 2. The posterior of model 1 was used to predict the values by session (a), and their dependence on MDT scores (b). Evidence ratios for the difference of sessions 5 and 7 from session 1 (m0) were strong (132.3, 375.0), as were those with respect to session 2 (m3). Sessions 5–7 did not differ, nor 1–3. Note that for each session, the other predictors are set to their relevant mean or modal values, while the MDT graph shows its influence at a single modal position of the other predictors and generates the full range of outcome values. For 2b, there is a strong evidence ratio for the positive slope (28), the same as for the coefficient in the model, since MDT score is not involved in interactions of group effects in the model.

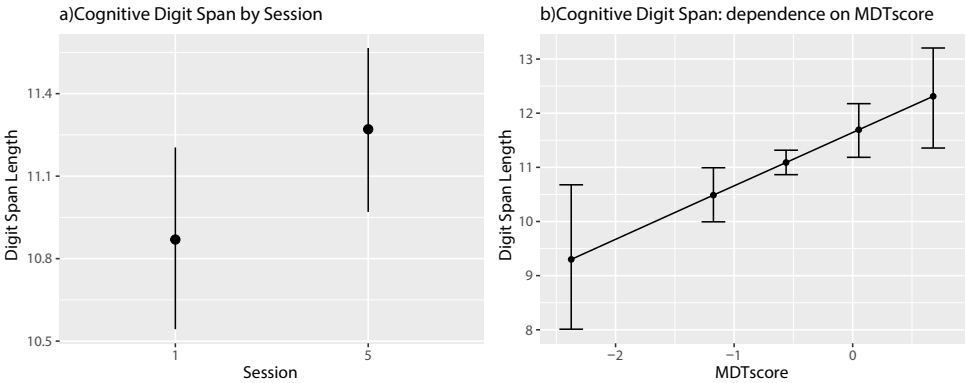

Figure 3. The DST results (a) per session (baseline and session 5 at the m12 point), and (b) per MDT score. The full range of outcomes was predicted by the MDT scores, with a strong e.r. (160.3) for the dependence. While the improvement between the m0 and m12 sessions was visible, its e.r. was weak.

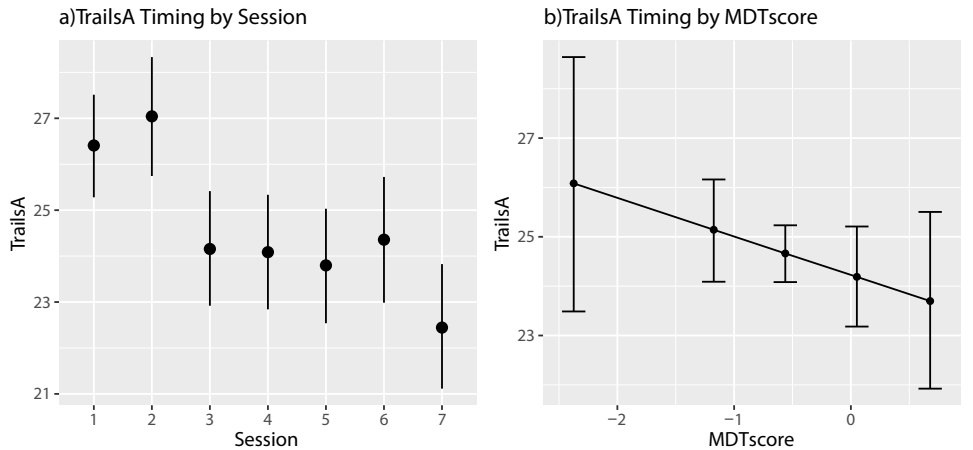

Figure 4. TMT A timings (seconds) are shown in relation to (a) session number and (b) MDT score. The e.r. for decreases in timing from Session 1 (m0) is strong for Sessions 3–7 (m3–m18: respectively, 205.9, 332.2, 1199.0, 217.2, Infinite, which latter can be interpreted conservatively as greater than or equal to 361, the number of data points). Most of the range of values can be explained by the dependence on MDT score (b), though the e.r. for the gradient of this response is weak (6.2).

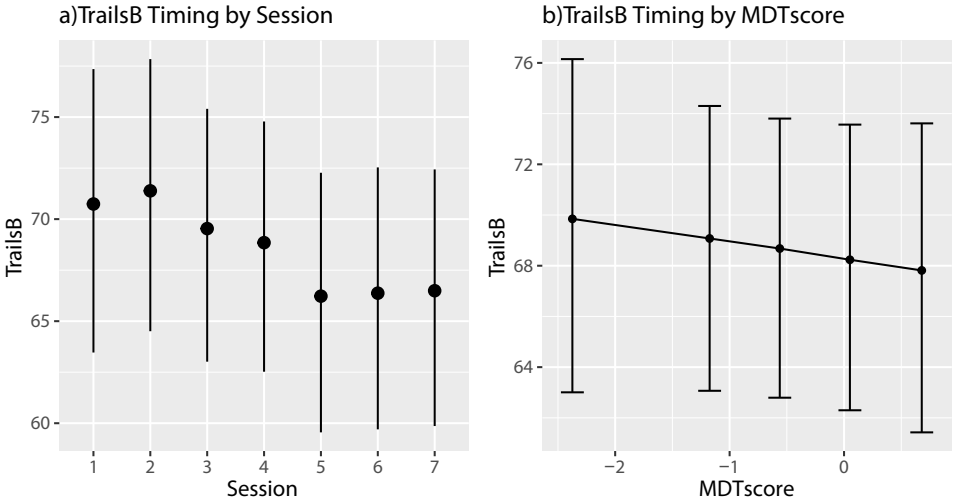

Figure 5. TMT B timings (seconds) are shown in relation to (a) Session number and (b) MDT score. The by-session graph (a) gives strong evidence ratios for improved timing in comparison with session 1 from sessions 4 to 7 (respectively, reporting 84.7, 1499, Inf, Inf). Session 7 was better than 5 with an e.r. of 78.0.

(by a count of 1.27, mean count was 9.85, e.r. 589.9) than after one block of replication training (or than in the zero training control, e.r. 41.7).

Another strongly evidenced feature of the model was a negative coefficient of −0.29 on age (e.r. 1167) suggesting that in this group older age was related to less diverse conception of alternative uses for objects: for example, for those aged 65–71, the mean was approximately 11, and for those aged 72–79, the mean was approximately 8. This resulted in a model difference of 2.38 between the two groups (strong e.r. infinite). Finally, the lowest three categories of subjective cognitive decline gave higher performance (by a mean of 1.77 counts) compared with the higher two (e.r. 3249). Given that it is difficult to assess the potential range of uses for each individual object, and hence to normalize the data, such observations should always be treated with caution.

3.3. Self-report measures

As there were no experimental manipulations or hypotheses made to directly assess the impact on health and well-being, the self-report measures comprising those on health and well-being and self-efficacy were assessed with simple descriptive models, including session as population effect and with group effects by session and participant. Each model provided a good fit, for example, with Bayesian R^2 in all cases greater than or equal to 0.69.

3.3.1. QuickDASH

The DASH survey allows participants to indicate their level of difficulty with arm and shoulder movements in routine daily activities: for simplicity, we used the standard scoring system, amalgamating 11 measures in the QuickDASH questionnaire, where values range from 0 to 100, and lower values are better. Our participants mostly had low values across the whole year of study (lowest 0, median 9.09, maximum 70.46), and while the final group value was numerically lower (better) than the start value, there were no strongly evidenced changes.

3.3.2. Basic Psychological Needs Scale

BPNS and subsequent GSE data were assessed with simple descriptive models, including session as population effect, and with group effects by session and participant. The BPNS Autonomy scale (observed min 2.71, median 5.85, maximum 7.00) and its Relatedness scale (corresponding observed values 2.00, 5.86, 7.00) showed minimal changes across sessions, all very weakly evidenced.

In contrast, the Competence subscale (corresponding observed values 2.66, 5.33, 7.00) showed marked enhancements after the first training block and thereafter, with strong evidence ratios from

80.6 to 427.6, excepting session 6 (for which e.r. 16.6). Session 5–7 values were not strongly evidenced as different from each other, so the effects from whatever was learned seemed to be retained in the post-training period (figure 6).

3.3.3. General self-efficacy

Consistent with the results for BPNS, in the GSE assessment, General subscale (observed minimum 55.00, median 93.00, maximum 119), there was strong evidence that sessions 3, 6 and 7 showed a more positive response than session 1 (respective evidence ratios 63.17, 1199.00, 243.9), while sessions 5, 6 and 7 were not strongly evidenced as different, showing a maintenance of the enhanced self-assessed efficacy (figure 7).

Given that one of our hypotheses (H1) relates to the predicted impact of musical improvising tasks on a faster rate of musical learning in comparison with replication, it was relevant to assess the potential impact of training task on the participants' reports of self-efficacy (figure 8). Evidence ratios were strong for the decline compared with start (0–0) values with one improvisation block at 1–0 (36.50), but not at 1–2. Positive effects of replication were strong at 1–1 (e.r. 32.9) but not at 0–1. Point 5 (1–2) was not strongly different from the preceding points, but 2–1 was strongly evidenced as enhanced. 2–2 amalgamates both data immediately following the last training block and that of the two post-study periods (over six months) and confirms a retention of the enhanced self-assessment (e.r. 1999); it also reveals a clear benefit for the second burst of improvisation training (e.r. for positive change versus 1–2, point 5, being 704.88). The retention of enhanced self-assessment in the post-study period is more fully detailed in an analysis by session, since this amalgamation is avoided.

For the GSE Social subscale, the representative observed values were 16.00, 28.00 and 42.00. The model showed sigma 2.68 and Bayesian R^2 0.73. In slight contrast to the BPNS 'relatedness' scale, here there were strongly evidenced enhancements at sessions 3–5 (e.r. 213.29, 74.95, 20.74), but responses returned to baseline in sessions 6/7 (three–six months after training ended) (figure 9).

3.3.4. Socio-cultural activity estimations

Given the evidence of influences of the training upon social involvement, we also assessed other potentially relevant participant variables recorded in the self-report data: of particular interest were extent of musical involvement, reading, TV watching and other cultural activities. The data were incomplete, in that out of a possible 391 values, we had between 291 and 322, which may have limited some analyses.[11] Nevertheless, there were no highly evidenced effects of session number (as sole predictor together with group effects) upon these responses.

4. Discussion

Our results show that in general, over the course of 12 months of musical instrument training, cognitively intact older adult novices experience gains in cognitive and motor skills, and that these can be predicted relatively successfully by gains in aural music skill measures. Our design permitted separation of the progressive effects of improvisation training, replication training, piano keyboard training, iPad keyboard training and duration of social exposure as a result of the training overall. In [13], we show a counterfactual model isolating the possible effects of passage of time in social and activity exposure: an effect was not detectable. Analogous models with the present data are in agreement with this conclusion and also confirm a lack of impact of the 2–7 repetitions of the cognitive and other tasks described on the level at which they were performed.

Our research also aimed to discern the effect of training task (replication and improvisation) and instrument (keyboard and iPad ThumbJam) on the cognitive tests. From anecdotal reports from participants and observations of the teacher, it is debatable whether the ThumbJam arrangement offers distinctly less motor 'load' on account of the flat surface as compared with the key depression required by the keyboard. For the purposes of this study, it was more fruitful to examine these as an opportunity to compare two different types of motor task for which models showed negligible differences.

The early impact of improvisation on the simple domain-general creative task, the AUT, was encouraging; however, the overall lack of effect of training on creative thinking as measured here is not entirely surprising, perhaps due to the limitations in the improvisation training (i.e. not entirely requiring creative thinking in a conceptual sense, but more being trained to use a certain toolbox of

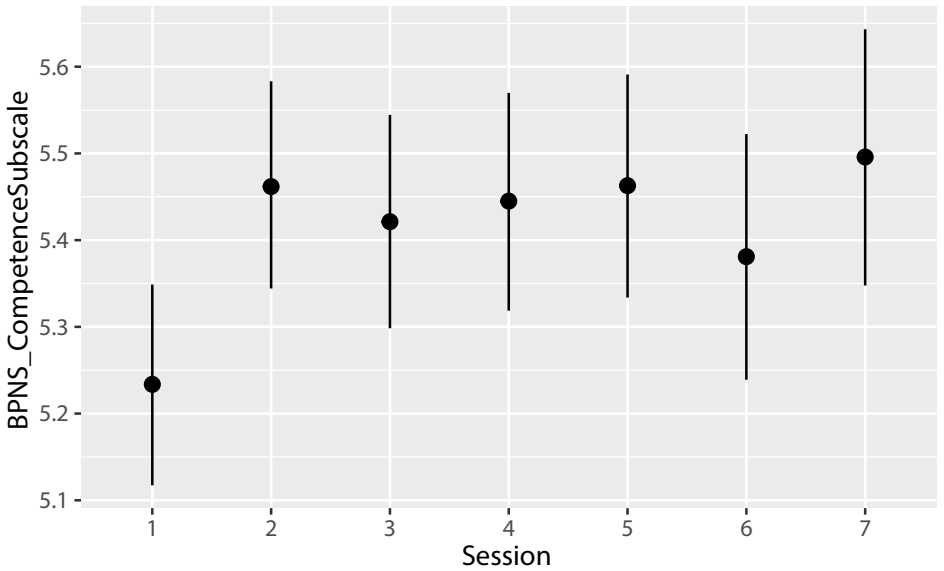

Figure 6. Participant competence, as measured on the BPNS scale, by session. This is an aggregate of multiple measures (each a 7-point scale within the 21-item BPNS questionnaire. Model Bayesian R^2: 0.69; sigma 0.48).

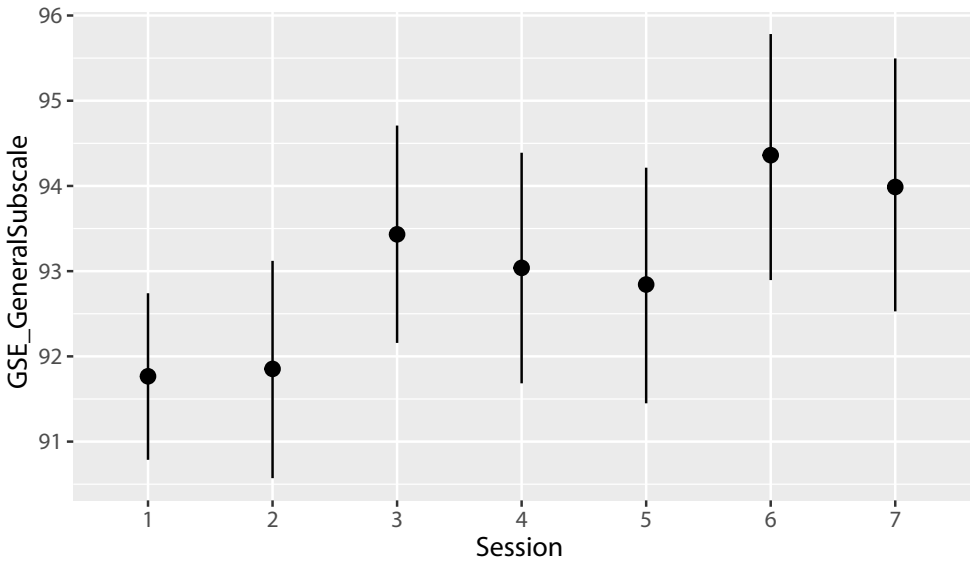

Figure 7. Improvement in response in the GSE general subscale. Sigma 5.62, Bayesian R^2 0.83.

melody manipulation techniques, or a creative 'reshuffling' of pitches). An alternative explanation is that learning to generate new musical material did not generalize to divergent thinking in the AUT image and text-based medium.

Our inclusion of health and well-being measures did not test any specific hypothesis on a direct impact of the experimental conditions, but was instead intended to serve as a useful gauge of the sample's holistic health and well-being over the testing period. Results suggest a steady maintenance of functional ability (from QuickDASH measures) and equally of basic psychological needs (after an early increase from the first bout of learning). The results concerning the impacts of training task on participants' self-efficacy are intriguing. These suggest that the first bout of improvisation represented a novel challenge for participants which made them partially question their levels of control in daily life activities, while this was not true for either bout of replication. It was clear from our interview data, as well as anecdotal reports from our teacher, that many participants faced the improvisation training with some trepidation, having previously accepted an image of music performance as being largely the replication of others' compositions. Improvisation tasks may also have created a higher

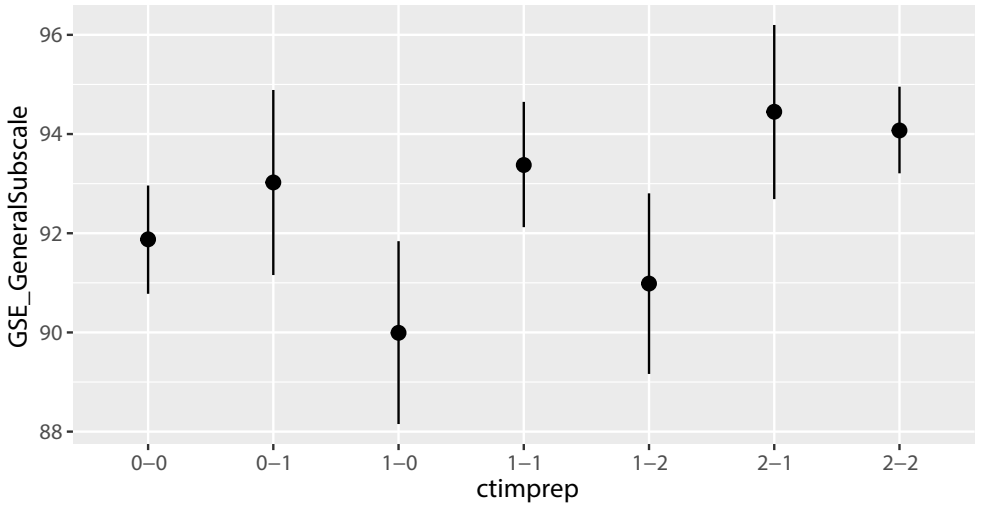

Figure 8. Improvement in GSE general control measures with respect to number of improvisation segments (the first number of the *X*-axis '*ctimprep*' terms) and replication segments (the second number). Predictions from the posterior.

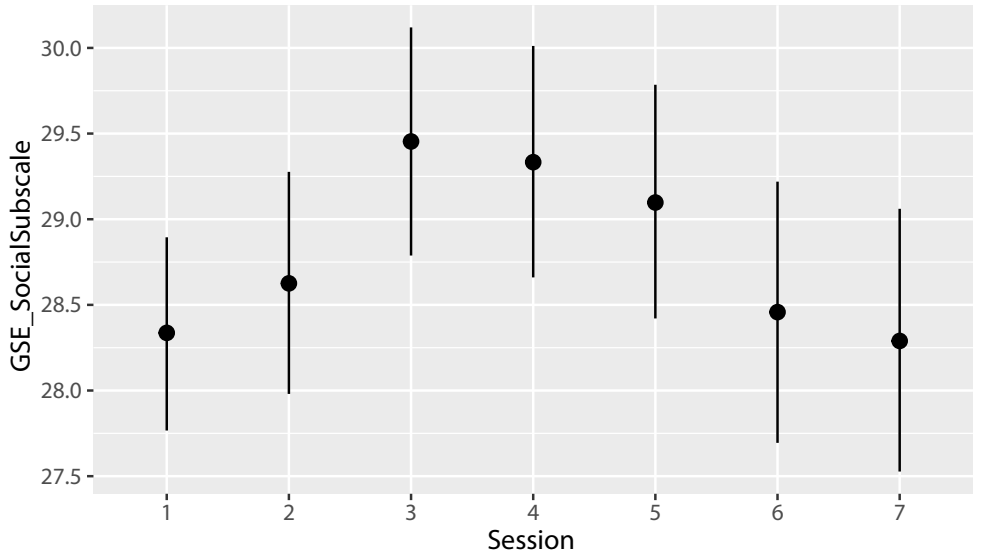

Figure 9. Transiently enhanced (sessions 3–5) GSE-measured social involvement during the training.

cognitive load compared with replication tasks due to the former's absence of a scaffold (most typically music notation as a support or guide, although in our case this was provided through aural-only recordings). This could well explain the influence of the first improvisation block on the GSE General score. Happily, both measures (*ctimprep* 2–1, 2–2) with 2 blocks of improvisation experience showed enhanced scores compared with the baseline, probably indicating that this trepidation was overcome with more improvisation experience and hence contributed to an increased awareness of the ability to foster control when required. This, alongside reference to the positive influence that exposure to improvisation training appears to play on the MDT results (and consequently how these appear to lead to general cognitive increases in executive function), may point towards investigations of larger-scale improvisation training where, for example, new structural features are established and manipulated.

Were our results due to the training programme or to cognitive advantages arising from pre-existing demographics and/or abilities [8]? Our crossover design and its analysis evidence the former, and specifically the learning achieved during it. The most convincing case for far transfer to executive function is in the TMT B results, with gains from this specific test having been reported across systematic reviews within the music training literature [6], and a case which arguably demonstrates that the impact from musical training has some domain-general transfer. Interrogating this result in terms of analysing the difference of test performance rather than absolute values gave a negligible

difference in result outcomes. Analysing the differences is a standard way of attempting to remove differences of individuals at the start of training. However, an issue with using the difference rather than absolute is that it removes a variance distribution of the first sample, hence reducing the overall information about variance: the essence of a Bayesian/Gelman approach for making what are latent variables in the SEM terminology (combining several measures putatively of one thing into a single one) is to use all the variance to get the best estimate of what is variance and what is signal (as we discussed previously [13]). Item analysis (which uses a kind of variance spreading by nesting the groupings—see formula in the electronic supplementary material), and the latent musical score values, tend to strengthen the interpretability of the observed differences (enhance the evidence ratio, lower the *p*-value for inferential tests). Consistent with the limited evidence of systematic effects of age or SCD on the acquisition or the cognitive and motor abilities assessed above, in our preceding analysis on the music performance data [13] we also found that these predictors were not commonly powerful in relation to the acquisition of musical skill as judged by melodic discrimination ability itself. Another element of support would be in the comparison of music sophistication values (from the Gold-MSI) between our sample and the general population (see [13, electronic supplementary material] indicating that our participants were in all respects musical novices). We can be reasonably confident, therefore, that our design and analysis are rigorous enough to withstand any other variability.

The relationship observed between trained music abilities (MDT) and measures of working memory (Digit Span Test) may only be relevant in the auditory domain. Our music training activities, particularly as they featured an aural-only method of delivery instead of relying on traditional visual music notation, may have served to improve discrimination abilities, perhaps also mediated by levels of attentional control in the auditory domain [52]. This could explain the strong relationship between the MDT scores and the Digit Span Test, although the current data are limited in only considering the forward version of this test, and only at baseline and m12 time points. Further investigation would be needed to clarify if this effect would be observed across tests of working memory in other modalities.

Learning in the musical domain better predicts the outcome of domain-general cognitive tests as compared with length of exposure to a music training intervention. Our results modelling each outcome measure both by session number and by MDT score demonstrate the advantage that progressive music learning has as a predictor over using a variable that simply describes exposure to a training programme. Schellenberg's main argument is that duration of training is often confounded with demographic, cognitive and personality variables and predicted by pre-existing musical ability [8]. Our design and analysis have been constructed to answer this as much as possible (we do not measure personality variables). So, would the same effect occur with training auditory skills via music-listening tasks rather than a multi-domain task (and expensive and time-intensive programme) such as melody replication and improvisation? Given the differences we observe between improvisation and replication training effects, this seems unlikely.

Limitations in our design were in restricting the training to uni-manual rather than bi-manual tasks in order to allow comparison across the two types of instrument. Pragmatic concerns over switching back and forth across instruments also prevented us from analysing every type of combination of task and instrument in sequences. Additionally, the move to online training and testing during the COVID-19 pandemic, although well received by participants, led to two withdrawals by participants who were not able to meet the extra technical challenges. The results we have attained should be generalized with caution, acknowledging that our sample was predominantly female and Australian nationality. The insensitivity of the screening tool mini-ACE III also means that our sample may comprise a greater range of individual differences. Our study crossover design allowed participants to experience all conditions of task and instrument training, with each acting as their own control. Instead of providing a between-participants comparison of how cognition might be affected by training versus an active or inactive control, our design and analysis procedures allowed us to model counterfactual conditions, separating the effect of passage of time and the social effect of lessons from the effects of the training itself. These counterfactual models showed negligible differences in the outcome measures and thus support the argument that the musical instrument training is responsible for effects observed.

This work has two main implications. The first is in considering how improvisation tasks are incorporated into music training designs. There are small positive signs from our analyses concerning impacts of early improvisation work. This is notable given the general lack of concentration on this particular task in music education or music for health programmes (where the majority is focused on replication tasks). Larger-scale training in improvisation methods beyond the limited scope we present here may have the potential to enhance both music learning and consequently the domain-general impacts of training music students of any age how to establish and manipulate new structural features.

The second related implication is in considering how we better structure musical interventions generally for older adults. Music in general is 'broad and undefined' [8] but we have taken care to describe the different facets of our training programme in detail [13,19] as well as the experiences in transitioning the training programme to an online format [50]. That is not to say that other music-performing tasks, contexts or instruments might not achieve positive results. Music training programmes for older adults have, in the majority, been designed and analysed based on an assumption that cognitive stimulation (or exposure to the challenges of learning a musical instrument) is the most important factor for any general cognitive gain, with scant attention given to the musical proficiency developed [13,53]. Von Bastian and colleagues' review of cognitive training literature [7] would suggest that our results show a potential far transfer, i.e. from music-based tasks (MDT) very similar to the musical instrument training, to untrained measures in contexts different to the music-based training (employing processing speed—TMT—and working memory capacity—DST); these could possibly be explained in terms of enhancing cognitive efficiency through improvements in probabilistic interference (both working memory and perception and attentional control), acquisition of strategies and more efficient attention allocation (working memory). Further investigation would be needed to draw firm conclusions in this regard. Music programmes for older adults should be considered both for their potential to increase cognitive capacity and for their potential to improve cognitive efficiency.

The implication of this in terms of designing such music training programmes for older adults, in line with our results, is that an individual's gain in musically relevant skills *does* matter. Thus, if the current structure or tasks included in music training do not benefit the individual, it is the training that should be personally tailored to ensure their success. This furthers discussion into widening participation for older adults in creative arts interventions and different ways training and involvement could be structured to meet individual differences. Attending to individual differences may see improvisation benefitting those individuals who find replication distinctly challenging [54], or alternative forms of notation being employed to meet difficulties with memory. Adjustments to the design of musical training would be to better target the cognitive challenge to the level of available resources the individual has to meet such challenges [55], providing further opportunity to examine tailored and potentially adaptive music programmes.

Ethics. All procedures performed in the studies involving participants were approved by the Human Research Ethics Committee of Western Sydney University (Ethics approval no.: H13206). Informed written consent was obtained from all participants for their participation and data use.

Data accessibility. The data for this paper are publicly available on the Open Science Framework [56]. Supplementary material is available online [57].

Declaration of AI use. We have not used AI-assisted technologies in creating this article.

Authors' contributions. J.M.: conceptualization, data curation, formal analysis, funding acquisition, methodology, project administration, writing—original draft, writing—review and editing; A.C.: data curation, formal analysis, methodology, writing—review and editing; C.J.S.: conceptualization, funding acquisition, methodology, writing—review and editing; R.T.D.: conceptualization, data curation, formal analysis, funding acquisition, methodology, project administration, supervision, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This work was supported by the Australian Research Council, Discovery Project (DP190102012 to J.M., R.T.D. and C.J.S.) and UK Research and Innovation Future Leaders Fellowship (MR/T040580/1 to J.M.).

Acknowledgements. We (and the participants) hugely appreciated the engaging and encouraging teaching of Mr Patrick O'Donnell, who also gave excellent demonstrations of replication and improvising. Critical to the online classes was Dr Madeleine J. Cannings, our technical support person and assistant for the end-of-block Performance Sessions. Other crucial research assistance was provided by Ms Siyao Cheng and Dr John R. Taylor. We also acknowledge valuable inputs from Prof. Andrea Creech.

Endnotes

[1]ThumbJam has been reported in use with older adults in varying musical contexts in [20].

[2]One participant reported having more than 2 years experience, having performed casually in a choir for 6 years. However, this participant had no formal training and so was within the inclusion criteria. Baseline Melodic Discrimination Test (MDT) data for this participant were compared with the rest of the dataset. Baseline MDT scores for the dataset ranged from -2.88 to 0.79 (M = -0.067, s.d. = 0.92). This participant produced a score below average (-0.93) for MDT at baseline. We therefore conclude that they were not specially skilled and retained their data for full analysis. See [13, electronic supplementary material] for further discussion on participant musical background and analysis of MDT.

[3]This was initially monitored through written practice journals, but there was poor adherence to filling these out. Given the large battery of other tests included in our measures, we opted to monitor adherence to daily practice via the interviews.

[4]There is no consensus yet on what the optimal dosage of a piano training intervention is (see discussion of [22], although studies operate either on lessons only or a lesson plus practice model).

[5]A bespoke screening test was created in *Max/MSP* to assess participants' ability to perceive whether a pair of piano tones were ascending or descending. This was included primarily to identify anyone who might need additional training to support perception of pitch direction and its translation to a horizontal surface to the teacher. The parameters of the pitch direction test are detailed further in (Chmiel *et al*. [23]). Participants were presented with piano tones ranging from one to four semitones and asked to judge whether the second tone was higher or lower in pitch. From 10 trials for each participant (including two 'practice trials'), 584 out of 680 trials were accurately answered (86%). Forty participants (59%) made at least one error from the 10 trials, with the maximum number of errors being 6 out of 10 trials incorrect. Further work analysing the results of this screening test and how this aspect of pitch perception might be impacted via targeted interventions is under way, although the test itself is not yet validated [23].

[6]It should be noted that the mini-ACE has questionable reliability for the diagnosis of dementia and is only recommended as an adjunct to a fuller clinical assessment (see [27]). For the purposes of our research study, we used this mainly as a screening tool, but also were asking participants about their perception of any subjective cognitive decline (SCD).

[7]Data are not available for this question for the first group of participants (n = 7).

[8]Group 1 participants (n = 7) completed the nine-hole pegboard test [28] to assess fine motor function. This test was abandoned once the project moved online as there was no satisfactory online equivalent at the time.

[9]https://neura.edu.au/resources/content/Trails-iPad_validation.pdf.

[10]Preliminary tests were conducted prior to analyses to examine the accuracy of the manual recording method. Ten cases where the tapping had been correctly captured by the app were compared with the video capture method. In each case, both capture methods produced the exact same number of taps, supporting the accuracy of this back-up recording method.

[11]Due to the high volume of self-report questions it was decided to make these socio-cultural activity questions optional. Some participants chose not to answer, and the first group of participants did not record data for these questions.

References

1. Kraus N, White-Schwoch T. 2014 Music training: lifelong investment to protect the brain from aging and hearing loss. *Acoust. Aust.* **42**, 117–123.
2. Hanna-Pladdy B, MacKay A. 2011 The relation between instrumental musical activity and cognitive aging. *Neuropsychology* **25**, 378–386. (doi:10.1037/a0021895)
3. Tymoszuk U, Perkins R, Spiro N, Williamon A, Fancourt D. 2020 Longitudinal associations between short-term, repeated, and sustained arts engagement and well-being outcomes in older adults. *J. Gerontol. B Psychol. Sci. Soc. Sci.* **75**, 1609–1619. (doi:10.1093/geronb/gbz085)
4. MacRitchie J, Breaden M, Milne AJ, McIntyre S. 2020 Cognitive, motor and social factors of music instrument training programs for older adults' improved wellbeing. *Front. Psychol*. **10**, 2868. (doi:10.3389/fpsyg.2019.02868)
5. McQuade L, O'Sullivan R. 2023 Examining arts and creativity in later life and its impact on older people's health and wellbeing: a systematic review of the evidence. *Perspect. Public Health* **144**, 344–353. (doi:10.1177/17579139231157533)
6. Rogers F, Metzler-Baddeley C. 2024 The effects of musical instrument training on fluid intelligence and executive functions in healthy older adults: a systematic review and meta-analysis. *Brain Cogn*. **175**, 106137. (doi:10.1016/j.bandc.2024.106137)
7. von Bastian CC, Belleville S, Udale RC, Reinhartz A, Essounni M, Strobach T. 2022 Mechanisms underlying training-induced cognitive change. *Nat. Rev. Psychol*. **1**, 30–41. (doi:10.1038/s44159-021-00001-3)
8. Schellenberg EG, Lima CF. 2024 Music training and nonmusical abilities. *Annu. Rev. Psychol*. **75**, 87–128. (doi:10.1146/annurev-psych-032323-051354)
9. Worschech F, James CE, Jünemann K, Sinke C, Krüger THC, Scholz DS, Kliegel M, Marie D, Altenmüller E. 2023 Fine motor control improves in older adults after 1 year of piano lessons: analysis of individual development and its coupling with cognition and brain structure. *Eur. J. Neurosci*. **57**, 2040–2061. (doi:10.1111/ejn.16031)
10. Mack M, Marie D, Worschech F, Krüger THC, Sinke C, Altenmüller E, James CE, Kliegel M. 2024 Effects of a 1-year piano intervention on cognitive flexibility in older adults. *Psychol. Aging* **40**, 218–235. (doi:10.1037/pag0000871)
11. Harrison PMC, Collins T, Müllensiefen D. 2017 Applying modern psychometric techniques to melodic discrimination testing: item response theory, computerised adaptive testing, and automatic item generation. *Sci. Rep*. **7**, 3618. (doi:10.1038/s41598-017-03586-z)
12. Harrison PMC, Müllensiefen D. 2018 Development and validation of the computerised adaptive beat alignment test (CA-BAT). *Sci. Rep*. **8**, 12395. (doi:10.1038/s41598-018-30318-8)
13. Chmiel A, Dean RT, Stevens CJ, MacRitchie J. 2025 Objective demonstration and quantitation of musical learning in older adult novices across a 12-month online study. *PLoS One* **20**, e0320055. (doi:10.1371/journal.pone.0320055)
14. Azzara CD. 1993 Audiation-based improvisation techniques and elementary instrumental students' music achievement. *J. Res. Music Educ.* **41**, 328–342. (doi:10.2307/3345508)
15. Zhang JD, Schubert E, McPherson GE. 2020 Aspects of music performance that are most highly related to musical sophistication. *Psychomusicology* **30**, 64–71. (doi:10.1037/pmu0000252)
16. Bugos JA, Gbadamosi A, Laesker D, Chow R, Sirocchi S, Norgaard M, Ghent J, Alain C. 2024 Jazz piano training modulates neural oscillations and executive functions in older adults: a pilot study. *Music Percept.* **41**, 378–392. (doi:10.1525/mp.2024.41.5.378)

17. Connell A. 2025 Music instrument learning throughout ageing and cognitive impairment. PhD thesis, Western Sydney University, Sydney, Australia.
18. Kakihara M, Wang X, Iwasaki S, Soshi T, Yamashita M, Sekiyama K. 2025 The association between music performance skills and cognitive improvement in a musical instrument training program for older adults. *Psychol. Music* **53**, 397–413. (doi:10.1177/03057356241248086)
19. Dean RT, Chmiel A, Radnan M, Taylor JR, MacRitchie J. 2022 AMMRI: a computational assessment tool for music novices' replication and improvisation tasks. *J. New Music Res*. **51**, 262–277. (doi:10.1080/09298215.2023.2270973)
20. Creech A, Varvarigou M, Hallam S. 2020 Developing musical possible selves through learning with technology and social media. In *Contexts for music learning and participation: developing and sustaining musical possible selves*, pp. 223–238. Cham, Switzerland: Palgrave Macmillan. (doi:10.1007/978-3-030-48262-6_12)
21. 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**, e89642. (doi:10.1371/journal.pone.0089642)
22. Lister JJ, Hudak EM, Andel R, Edwards JD. 2023 The effects of piano training on auditory processing, cognition, and everyday function. *J. Cogn. Enhanc*. **7**, 97–111. (doi:10.1007/s41465-023-00256-z)
23. Chmiel A, Taylor JR, Dean RT, MacRitchie J. In preparation. Training pitch discrimination and spatial pitch associations in music novices
24. Hsieh S, McGrory S, Leslie F, Dawson K, Ahmed S, Butler CR, Rowe JB, Mioshi E, Hodges JR. 2015 The Mini-Addenbrooke's Cognitive Examination: a new assessment tool for dementia. *Dement. Geriatr. Cogn. Disord*. **39**, 1–11. (doi:10.1159/000366040)
25. Matías-Guiu JA, Valles-Salgado M, Rognoni T, Hamre-Gil F, Moreno-Ramos T, Matías-Guiu J. 2017 Comparative diagnostic accuracy of the ACE-III, MIS, MMSE, MoCA, and RUDAS for screening of Alzheimer disease. *Dement. Geriatr. Cogn. Disord*. **43**, 237–246. (doi:10.1159/000469658)
26. Grasina A, Espirito-Santo H, Lemos L, Vilar MM, Simões-Cunha L, Daniel F. 2024 Mini-ACE: validation study among older people in long-term care. *J. Cogn*. **7**, 5. (doi:10.5334/joc.330)
27. Beishon LC, Batterham AP, Quinn TJ, Nelson CP, Panerai RB, Robinson T, Haunton VJ. 2019 Addenbrooke's Cognitive Examination III (ACE-III) and mini-ACE for the detection of dementia and mild cognitive impairment. *Cochrane Database Syst. Rev.* **12**, CD013282. (doi:10.1002/14651858.CD013282.pub2)
28. Earhart GM, Cavanaugh JT, Ellis T, Ford MP, Foreman KB, Dibble L. 2011 The 9-hole PEG test of upper extremity function: average values, test-retest reliability, and factors contributing to performance in people with Parkinson disease. *J. Neurol. Phys. Ther.* **35**, 157–163. (doi:10.1097/NPT.0b013e318235da08)
29. Wechsler D. 2012 *Wechsler adult intelligence scale*, 4th edn. APA PsycTests. (doi:10.1037/t15169-000)
30. Gignac GE, Weiss LG. 2015 Digit span is (mostly) related linearly to general intelligence: every extra bit of span counts. *Psychol. Assess*. **27**, 1312–1323. (doi:10.1037/pas0000105)
31. Bowie CR, Harvey PD. 2006 Administration and interpretation of the trail making test. *Nat. Protoc*. **1**, 2277–2281. (doi:10.1038/nprot.2006.390)
32. Bugos JA, Cooper P. 2019 The effects of mallet training on self-efficacy and processing speed in beginning adult musicians. *Res. Perspect. Music Educ*. **20**, 21–32.
33. Sánchez-Cubillo I, Periáñez JA, Adrover-Roig D, Rodríguez-Sánchez JM, Ríos-Lago M, Tirapu J, Barceló F. 2009 Construct validity of the trail making test: role of task-switching, working memory, inhibition/interference control, and visuomotor abilities. *J. Int. Neuropsychol. Soc*. **15**, 438–450. (doi:10.1017/s1355617709090626)
34. Park SY, Schott N. 2022 The trail-making-test: comparison between paper-and-pencil and computerized versions in young and healthy older adults. *Appl. Neuropsychol*. **29**, 1208–1220. (doi:10.1080/23279095.2020.1864374)
35. Guilford JP. 1967 *The nature of human intelligence*. New York, NY, USA: McGraw-Hill.
36. Vartanian O, Beatty EL, Smith I, Forbes S, Rice E, Crocker J. 2019 Measurement matters: the relationship between methods of scoring the Alternate Uses Task and brain activation. *Curr. Opin. Behav. Sci*. **27**, 109–115. (doi:10.1016/j.cobeha.2018.10.012)
37. Reiter-Palmon R, Forthmann B, Barbot B. 2019 Scoring divergent thinking tests: a review and systematic framework. *Psychol. Aesthet. Creat. Arts* **13**, 144–152. (doi:10.1037/aca0000227)
38. Kennedy CA, Beaton DE, Smith P, Van Eerd D, Tang K, Inrig T, Hogg-Johnson S, Linton D, Couban R. 2013 Measurement properties of the *Quick*DASH (Disabilities of the Arm, Shoulder and Hand) outcome measure and cross-cultural adaptations of the *Quick*DASH: a systematic review. *Qual. Life Res*. **22**, 2509–2547. (doi:10.1007/s11136-013-0362-4)
39. Turchin DC, Beaton DE, Richards RR. 1998 Validity of observer-based aggregate scoring systems as descriptors of elbow pain, function, and disability. *J. Bone Jt Surg*. **80**, 154–162. (doi:10.2106/00004623-199802000-00002)
40. Sherer M, Maddux JE, Mercandante B, Prentice-Dunn S, Jacobs B, Rogers RW. 1982 The self-efficacy scale: construction and validation. *Psychol. Rep*. **51**, 663–671. (doi:10.2466/pr0.1982.51.2.663)
41. Deci EL, Ryan RM. 2000 The 'what' and 'why' of goal pursuits: human needs and the self-determination of behavior. *Psychol. Inq.* **11**, 227–268. (doi:10.1207/S15327965PLI1104_01)
42. Gagné M. 2003 The role of autonomy support and autonomy orientation in prosocial behavior engagement. *Motiv. Emot.* **27**, 199–223. (doi:10.1023/A:1025007614869)
43. Krause AE, North AC, Davidson JW. 2019 Using self-determination theory to examine musical participation and well-being. *Front. Psychol.* **10**, 1–12. (doi:10.3389/fpsyg.2019.00405)
44. Hallam S, Creech A, Varvarigou M, McQueen H, Gaunt H. 2014 Does active engagement in community music support the well-being of older people? *Arts Health* **6**, 101–116. (doi:10.1080/17533015.2013.809369)

45. Harrison PM. 2020 psychTestR: an R package for designing and conducting behavioural psychological experiments. *J. Open Source Softw.* **5**, 2088. (doi:10.21105/joss.02088)
46. Chmiel A. 2020 *Guide to creating a server for online R experiments using psychTestR*. Sydney, Australia: Western Sydney University. (doi:10.13140/RG.2.2.18849.43360)
47. Halstead WC. 1947 *Brain and intelligence; a quantitative study of the frontal lobes*, p. 206, vol. **xiii**. Chicago, IL, USA: University of Chicago Press.
48. Reitan RM, Wolfson D. 1993 *The Halstead-Reitan neuropsychological test battery: theory and clinical interpretation*, 2nd edn. Tucson, AZ, USA: Neuropsychology Press.
49. Seinfeld S, Figueroa H, Ortiz-Gil J, Sanchez-Vives MV. 2013 Effects of music learning and piano practice on cognitive function, mood and quality of life in older adults. *Front. Psychol.* **4**, 810. (doi:10.3389/fpsyg.2013.00810)
50. MacRitchie J, Chmiel A, Radnan M, Taylor JR, Dean RT. 2023 Going online: successes and challenges in delivering group music instrument and aural learning for older adult novices during the COVID-19 pandemic. *Music. Sci.* **27**, 596–615. (doi:10.1177/10298649221097953)
51. Marsman M, Wagenmakers EJ. 2017 Three insights from a Bayesian interpretation of the one-sided *P* value. *Educ. Psychol. Meas.* **77**, 529–539. (doi:10.1177/0013164416669201)
52. Troche SJ, Wagner FL, Voelke AE, Roebers CM, Rammsayer TH. 2014 Individual differences in working memory capacity explain the relationship between general discrimination ability and psychometric intelligence. *Intelligence* **44**, 40–50. (doi:10.1016/j.intell.2014.02.009)
53. Laes T, Schmidt P. 2022 Promoting a musical lifecourse towards sustainable ageing: a call for policy congruence. *Int. J. Community Music* **14**, 103–119. (doi:10.1386/ijcm_00040_1)
54. Dean R, Smith H. 2013 *Improvisation hypermedia and the arts since 1945*. London, UK: Routledge.
55. MacRitchie J, Garrido S. 2019 Ageing and the orchestra: self-efficacy and engagement in community music-making. *Psychol. Music* **47**, 902–916. (doi:10.1177/0305735619854531)
56. MacRitchie J, Chmiel A, Dean RT, Stevens CJ. 2025 Cognitive Changes on Music Learning in Older People. OSF (doi:10.17605/OSF.IO/U24DM)
57. MacRitchie J, Chmiel A, Stevens CJ, Dean R. 2025. Supplementary Material from: Progressively Learned Musical Ability Predicts Cognitive Transfer in Older Adult Novices: A Twelve-Month Music Instrument Training Program. FigShare. (doi:10.6084/m9.figshare.c.8184491)