3 Methodology

3.1 Study Design

For participant recruitment, newspaper articles and posters were created and published in the two study locations (Hanover and Geneva). From various responses, 72 healthy retirees were selected, with 32 from Geneva and 40 from Hanover. All participants were native or fluent in either German or French, depending on the site. The individuals were right-handed and had no more than six months of musical practice throughout their life. None of the participants had impaired or uncorrected auditory or visual acuity. If they had any current or past neurological diseases, mild cognitive impairments, or early-stage dementia, they were excluded from the experiment. Additionally, none of the participants had any cardiovascular disease, hypertension, obesity, diabetes, or clinical depression.

Tests

Initially, information regarding age, gender, income, and educational level was collected, as shown in Table 3.1. The income levels were defined on a scale of 1 to 5, indicating the percentage of the national average income: 1 (<25%), 2 (25-75%), 3 (75-125%), 4 (125-175%), and 5 (>175%). Similarly, educational levels were scored from 1 to 6, representing different levels of education: 1 (elementary school), 2 (middle school), 3 (high school), 4 (bachelor's degree), 5 (master's degree), and 6 (PhD), with higher scores indicating a higher socioeconomic status. The participants' cognitive performance was assessed using the Cognitive Telephone Screening Instrument (CogTel). CogTel consists of six subtests that evaluate prospective memory, verbal short-term memory, working memory, verbal fluency, inductive reasoning, and verbal long-term memory. The scores range from 0 (lowest) to 60 (highest) [53]. Additionally, the participants completed the Cognitive Reserve Index questionnaire (CRIq) to evaluate their cognitive reserve. The CRIq calculates a score based on their educational background, working experience, and the frequency of engaging in various leisure activities such as sports, culture, and travel. An average score is derived from these factors. Scores below 70 indicate a very low cognitive reserve, while scores above 130 indicate a high cognitive reserve [**?**].

To assess the participants' musical engagement ability, they were required to complete the Goldsmiths Musical Sophistication Index (Gold-MSI) self-report inventory [54]. The Gold-MSI consists of six sub-scales that measure different facets of musical sophistication: The Active Engagement evaluates active musical engagement behaviors such as reading about

Table 3.1: Demographic Information of the Sample

		mean(SD)
Age		69.6 (3.2)
Education		3.9 (1.4)
CogTel		31.0 (7.2)
Income		2.9 (1.0)
Site	Hannover	40 (55.6)
	Geneva	32 (44.4)
Sex	Male	41 (56.9)
	Female	31 (43.1)
CRIq		137.4 (15.1)

music and the time and money spent on musical activities. The Perceptual Abilities are musical abilities related to perception, such as listening skills. Musical Training measures the extent of musical practice and training. Singing Abilities reflects the skills and activities related to singing. The Emotional Response to Music summarizes the participant's emotional response to music. And the General Musical Sophistication score incorporates various aspects from the other sub-scales to provide an overall measure of musical sophistication.

As Table 3.2 shows the scores of the sample is way lower than the norm data provided by the developer. That fits the prerequisite that the participants shouldn't be musically acquaintanced beforehand. The Gold-MSI questionnaire, along with the Cognitive Telephone Screening Instrument (CogTel), was repeated after the intervention. The participants' moti-

Table 3.2: GoldMSI Scores of the Sample before the Intervention

	Mean(SD)	norm
Active Engagement	29.4 (8.3)	41.52 (10.36)
Perceptual Abilities	38.7 (9.5)	50.2 (7.86)
Musical Training	10.2 (3.2)	26.52(11.44)
Singing Abilities	26.9 (6.5)	31.67 (8.72)
Emotions	22.2 (6.0)	34.66 (5.04)
General Musical Sophistication	49.3 (12.1)	81.58 (20.62)

The norm data is taken from the large online survey "How Musical Are You?" by BBC LabUK and decribes data from 147,633 participants [15].

vation to learn the piano was also assessed by collecting data on their average daily practice time at home. This information was collected at three-month intervals, specifically after

Table 3.3: Timeline and Tests

time	activity
t0/0 months of intervention	
	CogTel
	GoldMSI
	CRIq
3 months of intervention	Ode to joy Simple
	Homework
t1/6 months of intervention	
	Homework
t2/12 months of intervention	
	CogTel
	GoldMSI
	Ode to joy Simple
	Ode to joy Normal
	Homework
t3/6 months after intervention	
	Music Diary

three months, six months, and 12 months of the intervention. In the six months after the intervention the participants were asked to note their average pratice time in a music diary.

Intervention

After collecting the baseline data, the intervention began, consisting of one year of piano practice for the participants. The practice sessions were conducted in pairs with a professional piano teacher and lasted for one hour per week. The lessons followed a pre-established curriculum (see A.2). In the classroom, three Yamaha keyboards were set up for the participants to use. Additionally, each participant was provided with a Yamaha keyboard to practice at home. The practice sessions started with imitation and listening exercises, allowing the participants to become familiar with the piano and adopt a relaxed and correct posture. Activities such as clapping, singing, and walking in rhythm were incorporated into the sessions to enhance rhythm and musicality. Throughout the intervention, the participants gradually learned to read sheet music. A method inspired by "Piano Prima Vista" by Jens Schlichting (Internote GmbH Musikverlag, 2013) and the Hal Leonard Adult Piano Method was utilized to teach the participants how to read and interpret musical notation. To reinforce learning and progress, the participants were encouraged to practice for approximately 30 minutes daily at home. They were given exercises and small pieces of music to practice, which they

would then present in the following week's session. Overall, the intervention aimed to provide structured piano practice, guided by professional teachers, and foster daily practice habits to enhance the participants' musical skills and enjoyment of playing the piano.

3.2 Evaluation of Piano Performance

Careful decisions were made to evaluate the piano performances, following the model described in chapter 2.1.3, in order to minimize possible sources of friction. The performance context and assessment's purpose were clear and unchanged throughout ratings, allowing for the exclusion of these factors from the analysis.

The evaluation of piano recordings was carried out by nine different raters, aged 20-30 ($M = 25\,78$), all with extensive experience in piano playing ($M = 18{,}1$ years). Six of the raters held at least a bachelor's degree in music studies, while two had a degree in music education and one a masters degree in psyhology. On average, the raters had three years of experience teaching pianos. Aligning the raters' background helped minimize potential differences in their ratings, although characteristic differences remained. These differences could be balanced in the statistical analysis. The raters were instructed to set aside their personal preferences and musical biases and to evaluate the recordings solely based on the given musical parameters. This ensured a more objective assessment and minimized the influence of subjective preferences on the evaluation results. All raters rated the recordings in different randomized orders to reduce possible effects of inattetion and dependency on their individual day form.

A scale ranging from 1 to 7 was used by the raters to evaluate six different musical aspects of the the recordings. A rating of 1 represented a very low rating, while 7 indicated a very high rating of the musical parameters. The use of a seven-point scale allowed the raters to discern and evaluate finer nuances and differences in the recordings.

The raters listened to the recordings twice. During the first pass, they focused on articulation, dynamics, and rhythm. The raters assessed the correct execution of indicated articulations in the musical score, such as staccato and legato. They also discerned differences in dynamics, such as forte and piano, as well as crescendos and decrescenos. Rhythmic accuray was evaluated based on whether the played notes occured in the correct temporal relationship, even if there were hesitations or interruptions. In the second pass, the raters evaluated fluency, pitch accuracy, and musical expressiveness. Since the subjects were completely new to the piano, fluency was rated very good when the performer could play the piece without interruptions. Pitch accuracy assessed playing only the correct notes. Musical expressiveness was intended to reflect the quality of phrasing and interpretation of the musical piece. By

providing these clear guidelines, the raters were able to assess the recordings based on a shared framework, thereby enabling a comparable evaluation.

To control the consistency of their ratings, the raters unknowingly evaluated 30 recordings twice. After each rater had submitted all evaluations, the results were statistically analyzed. The intraclass correlation coefficient (ICC) was initially calculated for each rater based on the double-rated recordings. The ICC serves as a measure of agreement between the ratings. A correlation of 0 indicates inconsistent ratings, with significantly different evaluations of the same recordings. A correlation of 1, on the other hand, indicates complete agreement in the evaluation of the double-rated recordings. The ICC(3,1) type, according to Shrout and Fleiss [55], was used, employing the two-way mixed effects model with a consistent relationship. The ICC was calculated by subtracting the variance between subjects from the residual variance and dividing the result by the variance between subjects. For the further analysis, the ratings of each rater were weighted by the respective ICC value per variable.

The use of a larger number of raters with a musical background and the application of an objective rating scale contributed to the objective assessment of the musical parameters of the piano recordings. This approach allows for an informed and comparable evaluation of the musical performance of the subjects and facilitated the development and improvement of their musical abilities.

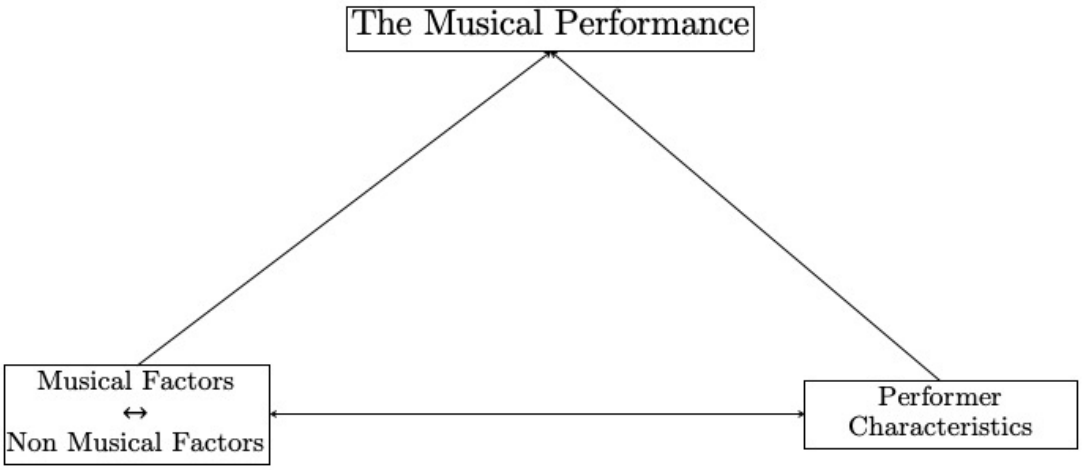


Figure 3.1: Impacts on Musical Performance

By minimizing the process model of McPhearson and Thompson [1] through careful disicions, the potential impact on the musical performances were reduced (see 3.1). Considerung per-

formance characteristics and their interactions with specific musical or non-musical factors could provide further insights and enhance the analysis of piano performances.

3.3 Evaluation of Musical Progress

To assess the progress of the participants' musical abilities, recordings were made after three and twelve months of piano lessons. After three months, the participants were introduced to a simplified version of Beethoven's "Ode to Joy" (refer to A.1) and given a two-week period to familiarize themselves with it. During the recording, the participants were encouraged to play continuously without restarting and to follow the instructions provided on the sheet music, including dynamics and articulation. This recording served as an evaluation of their progress after three months of piano practice. After twelve months of piano practice, the same version of "Ode to Joy" was recorded again to investigate the long-term effects of the piano lessons. Additionally, the participants learned a new and more challenging version of the same song, which was also recorded. Therefore, each participant had three recordings documenting their progress on the piano (see 3.2).

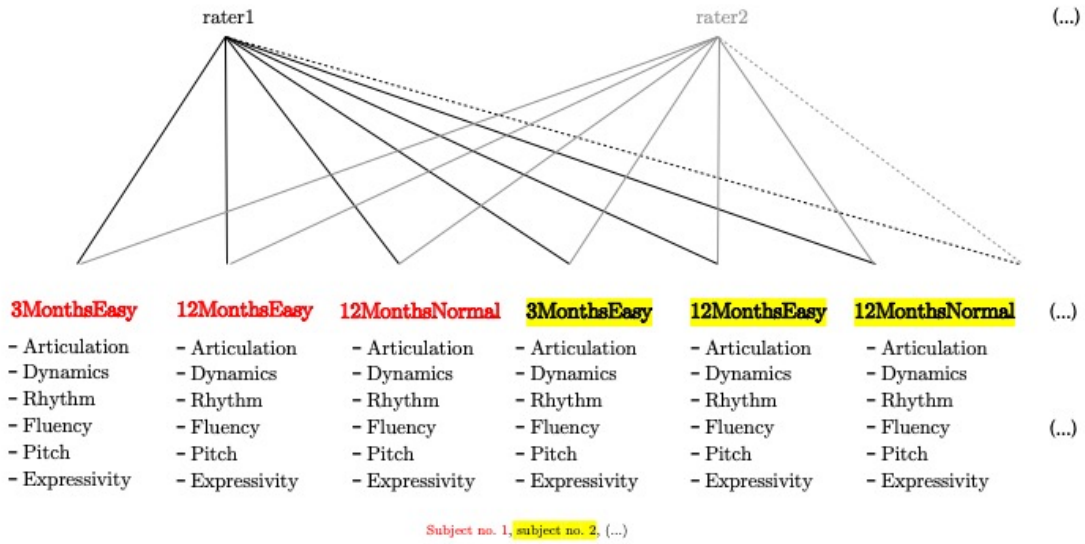


Figure 3.2: Rater System

The progress was evaluated by assessing changes in six variables: articulation, dynamics, rhythm, pitch, fluency, and expressivity over time in the easy version of "Ode to Joy." Possible predictors, such as age, sex, income, education, study location, CRIq score, and CogTel score, were included in the analysis to explain potential changes. The factors of the Goldsmiths

Musical Sophistication Index (GoldMSI) were also taken into account. Individual explanatory models were developed for each variable to identify predictors of improvement.

Furthermore, it was examined whether the predictors that indicated progress in the musical parameters of the easy version of "Ode to Joy" also predicted better results in the more challenging version of the song. The analysis aimed to identify individual explanatory models for each variable.

Having analyzed the six musical variables, the possibility of summarizing these variables into a more concise representation of *musicality* was explored. To achieve this, factor analysis was employed, a statistical technique that uncovers underlying latent factors that explain the observed correlations among the variables. By identifying one general *musicality* factor, we aim to capture the essence of participants' musical abilities and provide a comprehensive understanding of how these variables converge to form a cohesive musical aptitude. The generation of a general musicality factor through factor analysis will not only enhance the interpretability of the results but also allow us to draw meaningful conclusions about the participants' overall musical potential. By unifying the six variables into one overarching factor, we can more effectively examine the influence of demographic factors, such as age, gender, and musical training, on this comprehensive measure of musicality.

3.4 Statistics

The data analysis was conducted using a Bayesian mulitilevel model with the R package "brms", developed by Paul Bürkner [56, 57]. Bayesian statistics are well-suited for accommodating hierarchical models, especially when dealing with data from multiple sources or when there are dependencies among observations. In our study, data were collected from different groups (individuals) and included repeated measures from the same individuals (each rater rated all the participants), making a hierarchical model appropriate for capturing underlying variation and improving estimation accuracy. Bayesian apporaches have several advantages over frequentist statistics. Unlike frequentist methods that rely on statistical significance and produce dichotomous statements of significance or non-significance, Bayesian analysis provides a posterior distribution that encompasses all plausible values of the effect. Additionally, Bayesian statistics allow for the incorporation of prior knowledge or beliefs about the distribution of the data.

To analyze results based on a scale ranging from 1 to 7, we utilized Bayesian statistics with a beta distribution, which offers a flexible and intuitive framework for modeling data within a bounded range [58]. The beta distribution is characterized by two shape parameters, α and β, controling the distribution's shape and behavior. It is well-suited for modeling proportions, probabilities, or continuous variables that are constrained within a finite range,

such as Likert-scale ratings ranging from 1 to 7. The versatility of the beta distribution allow it to model various response distributions, including skewed, symmetric, or bimodal distributions, making it applicable to various types of data. By specifying a prior distribution for the beta parameters (α and β), prior information can be included in the analysis, which can be particularly useful when limited data are available. Bayesian inference provides a framework for quantifying uncertainty in parameter estimation. By combining prior beliefs with observed data, posterior distributions can be obtained, representing a range of plausible values for the parameters. This uncertainty estimation is often valuable when interpreting results and making decisions based on the data. Within this Bayesian framework, we reported 95% credible intervals (CI), indicating that there is a 95% probability for the effect to fall within this range. CIs that do not include zero suggest a high likelihood of a real effect, while CIs strongly overlapping with zero indicate no effect.

For each variable (articulation, dynamics, ryhthm , fleuncy, pitch, expressivitiy), a Bayesian multilevel approach was employed using the following regression equation:

$$Variable \sim Demographic * time + (1 + time|Code) + (1 + time|rater) \tag{3.1}$$

Prior to analysis, the variables and the possible demographic predictors were centered at their means and standardized. Dummy variables (0|1) were used to encode sex (female|male) and site (Hanover|Geneva). A one-unit change refers to a change of one standard deviation. Each variable was analyzed individually with respect to potential demographic predictors. The models allowed the slopes and intercepts of the participants to vary for a better model fit. The intercept represents the baseline performance at the beginning of the intervention, while the time effect reveals the change in performance over the study duration. The data were differentiated by the Code, which represents the participants' ID, and different rater effects were considered. The models convergered with a Rhat-value of 1.0, indicating a satisfactory convergence of the Markov Chain Monte Carlo algorithm used for estimation. To ensure a good fit, the different models were checked using the pp_check function.

In conclusion, the statistical methods in this study involved Baysian multilevel modeling with beta distributions for scale 1 to 7 data. This approach offered various advantages, including the flexibility to model bounded data, incorporation of prior knowledge, and quantification of uncertainty. The individual analysis of variables allowed for a comprehensive understanding of the effects and predictors of observed changes in musical abilities.

The following factor analyis aimed to uncover the complex relationships among the individual musical variables into a single latent factor, thereby revealing patterns of shared variance that contribute to participants' overall musical performance. This approach facilitates a deeper understanding of the underlying structure of musical abilities and highlight the fundamental dimension that underlies the participants' proficiency in various musical aspects. The data

used for the analysis was extracted from the calculated models, considering that the values are already weighted by the ICCs of the different raters. The number of latent factors was determined through scree plot analysis, a simple line plot that displays the eigenvalues of the factors in descending order against their corresponding factor numbers. The resulting factor loadings represent the strength and direction of the relationship between the observed variables and the latent factors. Higher absolute values indicate a stronger relationship between the observed variables and the latent factors. To calculate the *musicality* factor for each participant, the factor loadings ($FL_{variable}$) were multiplied by the rated values of the corresponding musical variables and then summed. The resulting composite measure represents the participant's overall *musicality* score, as shown in equation 3.2.

$$M = (FL_A \times A) + (FL_D \times D) + (FL_R \times R) + (FL_F \times F) + (FL_P \times P) + (FL_E \times E) \quad (3.2)$$

To assess the development of the potential *musicality* factor, three factor analyses were conducted. One was performed for the baseline and simple version of the "Ode to Joy" ($FA_{0,0}$), the second after the intervention and the simple version ($FA_{1,0}$), and the third after the intervention with the more challenging version of "Ode to Joy" ($FA_{1,1}$). To avoid complicating and potentially distorting the analysis, the underlying data was extracted using models that only calculated the intercepts of each of the three measuring points. Each analysis resulted in one *musicality* score for each participant at the beginning of the intervention playing the easy version of "Ode to Joy", after the intervention with the easy version, and after the intervention playing the more challenging version of "Ode to Joy".

Subsequently, Bayesian multilevel models were establised to explore potential demographic predictors. As the raters are already weighted in the input data and the input data are the estimate intercepts of all participants, the regression equation follows the simple struture shown in equation 3.3:

ja, stimmt das denn überhaupt? über die zeit gestreckt hat dann jeder Code n eigenen estimated slope oder?

$$Musicality \sim Demographic * time \quad (3.3)$$

Given that the data is not limited to specific values, the Baysian models are modeled following a normal distribution to capture the uncertainty and variability in the estimates. It is fully characterized by two parameters, the mean and the variance. The normal distribution's properties, such as being symmetric and unimodel, align well with the observed data.

4 Results

The evaluation of the piano recordings yielded promising results, offering valuable insights into the reliability and consistency of the rater system. The slopes and intercepts of the piano progress varied strongly among the participants, which is expected given their diverse individual backgrounds. The results are categorized into four main areas of evaluation: the rater system, the progress of piano playing, the progress in other musical abilities, and the exploration of a potential general *musicality* factor.

4.1 Rater System

The evaluation of the piano recordings involved nine raters who used a seven-point scale to assess the recordings based on six musical parameters: fluency, pitch, rhythm, articulation, expressivity, and dynamics. The scale allowed for the recognition and evaluation of subtle nuances and differences in the performances. To check the reliability and consistency of the evaluations, the raters rated 30 recordings twice. With these double rated evaluations the Intraclass Correlation Coefficient (ICC) values were computed to determine the level of agreement and consistency among the raters' evaluations. Table 4.1 presents the ICC values for each rater across the six variables. The ICC values indicated varying degrees of

Table 4.1: ICC Values of the Rater

rater	Fluency	Pitch	Rhythm	Articulation	Expressivity	Dynamics
1	.931	.733	.817	.818	.532	.884
2	.912	.737	.834	.873	.452	.834
3	.889	.81	.586	.64	.708	.486
4	NA	NA	.626	.716	NA	.378
5	.92	.886	.679	.716	.404	.647
6	.861	.91	.823	.848	.795	.473
7	.818	.795	.768	.797	.762	.729
8	.803	.783	.943	.858	.804	.765
9	.808	.88	.774	.8	.907	.915

> .9; excellent, .75 - .9; good , .5 - .75; moderate , > .5; poor reliability.

agreement among the raters for each variable. For fluency, the ICC values ranged from 0.803

to 0.931, indicating a high level of consistency among the raters' assessments. Similarly, pitch exhibited strong agreement among the raters, with ICC values ranging from 0.733 to 0.91. Rhythm evaluations also demonstrated a high level of agreement, with ICC values ranging from 0.586 to 0.943. Articulation was assessed with moderate to high reliability, as indicated by ICC values ranging from 0.64 to 0.858. Expressivity showed moderate agreement among the raters, with ICC values ranging from 0.404 to 0.907. The evaluations of dynamics exhibited the greatest variability among the raters, with ICC values ranging from 0.378 to 0.915. Overall, the results suggest that the raters demonstrated good to strong consistency in evaluating fluency, pitch, rhythm, and articulation. However, there was more variability in the assessments of expressivity and dynamics.

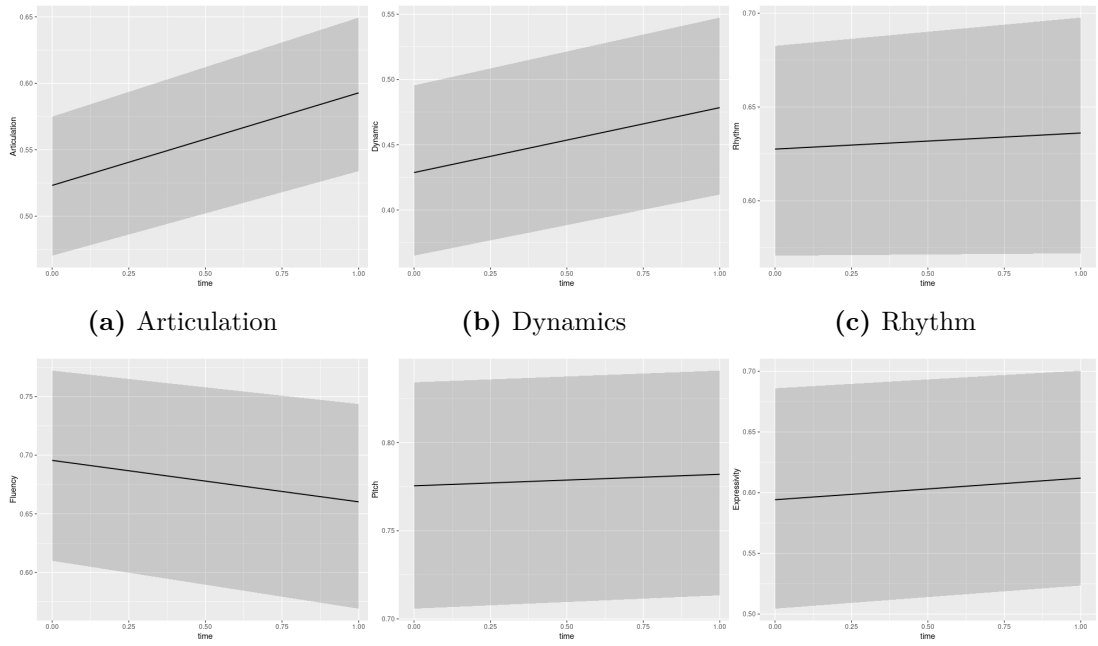
The ICC values reveal considerable variability in consistency among the raters. For instance, Rater 8 and Rater 9 demonstrate higher reliability compared to Rater 3 and Rater 4. Despite analyzing factors such as age, years of piano practice or teaching, no clear explanations for these differences were found, likely due to the small group size and alignment of raters in these factors. Furthermore, no differences were observed between different university degrees, such as music educators and musicians, in terms of rating. In the subsequent analysis, the ratings were weighted based on the ICC values per rater and variable. This approach ensured that raters with higher consistency had a greater impact on the analysis of musical performance compared to raters with lower ICC values.

4.2 Piano Progress

The analysis of piano playing progress across the six variables articulation, dynamics, rhythm, fluency, pitch, and expressivity, revealed strong variations in slopes and intercepts. Examining the changes in these variables over time, various patterns emerged (see Figure 4.1). Notably, while articulation showed the most improvement, the score for fluency actually decreased. Below, a detailed analysis of each variable is presented.

Articulation

The intercept for articulation is 0.523 (95% CI [0.47, 0.589]), indicating a moderate level of articulation proficiency at the beginning. Over time, there was a clear positive time effect (0.29, [0.08, 0.51]), suggesting that participants' articulation skills improved during the intervention. On average, the subjects displayed a 7% improvement in articulation as a result of the intervention. Although the effect values for demographic factors are relatively small, trends can be observed in relation to articulation scores. Participants with higher income (0.12, [-0.03, 0.27]) displayed higher articulation scores initially. Similarly, participants with higher cognitive reserve (0.13, [-0.02, 0.26]) exhibited higher articulation scores initially, both with a high probability of positive impact (96%). The goldMSI categories showed no effect



(a) Articulation **(b)** Dynamics **(c)** Rhythm

(d) Fluency **(e)** Pitch **(f)** Expressivity

Figure 4.1: Predicted Values over time

on participants' initial articulation performance. To understand if the effect of demographic factors on articulation changed over time, we examined the interaction between each demographic factor and time. While males showed more improvement over time (0.22, [-0.35, 0.20]), this effect was not relevant due to the overlapping credible interval with zero. Similarly, participants who spent more time on their homework displayed more improvement over time (0.11, [-0.07, 0.28]), with a 92% probability of being positive. The effect of musical engagement on articulation over time was clearly negative (-0.21, [-0.27, -0.06]). Participants with self-reported lower levels of musical engagement showed a clearer increase in articulation scores during the intervention. The effect of emotions towards music on articulation over time was also clearly negative (-0.16, [-0.31, 0.00]). Perceptual abilities, musical training, and singing abilities had a negative effect on articulation during the intervention, all being most likely to be negative. Participants with lower scores in these areas showed more increases in articulation compared to those with higher scores. General sophistication had a negative effect on articulation over time (-0.19, [-0.34, -0.05]). Participants with higher levels of general sophistication displayed a decrease in articulation scores during the intervention. After twelve months of intervention, participants with higher CogTel (0.1, [-0.05, 0.24]), higher cognitive reserve scores (0.1, [-0.06, 0.25]), and more perceptual abilities (0.1, [-0.5, 0.25]) tended to have higher articulation scores on the more difficult version of "Ode to Joy". Additionally, homework seemed to have a small impact (0.11, [-0.04, 0.26]).

Dynamics

The intercept for dynamic performance is 0.443 ([0.378, 0.52]), indicating a relatively low baseline performance. However, the time effect of 0.21 ([-0.01, 0.42]) suggests a positive trend over time, indicating that participants showed improvement in dynamics during the intervention. On average, the participants' dymanic skill improved by 5% over the course of the intervention. Initially, gender had an impact on the dynamic scores. Female participants scored generally better in the easy version of "Ode to Joy" in the first measurement (-0.25 [-0.52, 0.02]). Additionally, participants with higher CogTel scores showed higher dynamics results (0.1, [-0.04, 0.24]), with a 92% probability of remaining positive. Participants who self-reported having a high score regarding their emotions towards music scored clearly better than the others (0.14, [0.01, 0.28]). Over time, the younger participants improved more (-0.18, [-0.35, 0.00]). Unlike before, the males showed a slightly greater improvement (0.23, [-0.11, 0.58]). Again, participants with less musical emotion improved more (-0.27, [-0.42, -0.1]). A trend was observed regarding participants with less musical training, fewer singing abilities and lower general sophistication, as they showed more improvement than others. These effects have a probability of 87%. In the more difficult version of "Ode to Joy", participants with higher cognitive reserve scores performed better (0.19, [0.04, 0.33]). Participants with higher income, CogTel and more homework also achieved higher scores with a positive probabilty of 70%. Furthermore, participants with less emotions towards music scored higher scores on dynamics with a probability of 92%.

Rhythm

The intercept for rhythmic skills is 0.634 [0.523, 0.696], indicating a moderate baseline performance compared to other variables. The time effect is 0.04 [-0.13, 0.21], suggesting that rhythmic skills remained relatively stable over time, changing only by 0.2%. Regarding demographic factors, the analysis revealed that younger participants exhibited higher rhythmic skills at the beginning (-0.14, [-0.27, -0.01]), as did male participants (0.19, [-0.06, 0.45]). Additionally, participants with higher income showed higher rhythmic scores (0.14, [0.00, 0.28]). Cognitive reserve also had a positive effect on rhythmic skills (0.16, [0.03, 0.29]). However, no effect was observed for goldMSI categories. Over time, no demographic factors showed an effect on rhythmic skills. Instead, musical engagement had a negative effect (-0.17, [-0.32, -0.01]), suggesting an increase in rhythmic skills over time for participants with lower musical engagement. Similarly, emotions towards music also had a negative effect (-0.15, [-0.3, 0.00]). At difficulty level 1, participants from Hannover showed higher scores compared to Geneva (-0.24, [-0.52, 0.03]). Additionally, higher cognitive reserve was associated with better rhythmic scores (0.14, [0.00, 0.28]), as well as higher perceptual abilities (0.14, [0.00, 0.27]), and general sophistication (0.12, [-0.02, 0.26]).

Fluency

The intercept for fluency is 0.705 [0.622, 0.779], indicating a relatively high baseline performance. However, there is a negative time effect of -0.16 [-0.33, 0.02], suggesting a decrease over 4% in fluency over the intervention period. Regarding demographic factors, age showed a negative effect on fluency (-0.12, [-0.29, 0.05]), while being male had a positive effect (0.25, [-0.08, 0.58]). Both factors have a 92% porbability of being real. CogTel scores and being from Hannover also slightly impacted fluency (CogTel: 0.11, [-0.07, 0.28]; Hannover: -0.25, [-0.57, 0.08]). Higher cognitive reserve had a strong positive effect on fluency (0.23,[0.06, 0.4]), and no effect was observed for goldMSI categories. Over time, demographic factors showed varying effects on fluency. Participants from Geneva displayed an improvement over time (0.22, [-0.09, 0.53]), and also musical training had a positive effect (0.1, [-0.05, 0.26]). Emotions had a negative effect on fluency over time (-0.1, [-0.26, 0.05]). At difficulty level 1, being male had a positive effect on fluency (0.29, [-0.13, 0.7]). Also, being from Hannover had a positive effect (-0.23, [-0.64, 0.18]). Higher cognitive reserve was associated with improved fluency (0.21, [0.00, 0.42]), as well as higher perceptual abilities (0.24, [0.03, 0.44]), and singing abilities (0.15, [-0.06, 0.36]). Additionally, higher general sophistication showed a positive effect on fluency (0.2, [-0.02, 0.41]).

Pitch

The analysis of pitch data revealed a high baseline performance, with an intercept of 0.776 [0.709, 0.835]. Participants demonstrated a promising starting level of pitch accuracy. Over the intervention period, there was a almost no improvement in pitch skills, as evidenced by a time effect of 0.04 [-0.18, 0.27], indicating a very small positive trend (0.6%) over time. At baseline, Gender showed a negligible effect on pitch, with male participants having slightly lower scores, but the credible interval spanned from -0.49 to 0.15. On the other hand, CogTel scores and cognitive reserve had positive effects on pitch, with coefficients of 0.2 [0.04, 0.36] and 0.18 [0.02, 0.34], respectively. Improving the pitch performance education demonstrated a small negative effect, with an effect size of -0.1 [-0.25, 0.05]. Conversely, participants with lower CogTel scores improved more (-0.13, [-0.33, 0.08]). Participants from Geneva also developed their pitch skills a little more (0.18, [-0.21, 0.57]). Moreover, the amount of time spent on homework had a positive influence on pitch improvement (0.15, [-0.06, 0.36]). Cognitive reserve showed a negative effect, indicating that participants with higher cognitive reserve demonstrated less improvement in pitch accuracy (-0.22, [-0.43, -0.02]). Similarly, musical engagement was associated with reduced improvement (-0.25, [-0.44, -0.06]). Perceptual abilities and singing abilities also had negative effects on pitch improvement, with effect sizes of -0.18 [-0.38, 0.01] and -0.18 [-0.38, 0.02], respectively. Moreover, emotions towards music showed a strong negative effect on pitch improvement (-0.33, [-0.51, -0.14]). Lastly, participants with less general sophistication improved more in pitch performance (-0.2, [-0.39,

0.00]). At difficulty level 1, age showed a negative effect (-0.15, [-0.33, 0.02]), indicating that younger participants achieved higher pitch scores. Male participants continued to exhibit a positive effect (0.19, [-0.15, 0.54]). Similarly, CogTel scores had a positive effect (0.11, [-0.07, 0.29]). Participants from Hannover scored higher at the more challenging version of "Ode to Joy" (-0.25, [-0.59, 0.1]). Lastly, cognitive reserve had a positive effect, with an effect size of 0.16 [-0.02, 0.33], indicating that participants with higher cognitive reserve scores performed better on pitch evaluation.

Expressivity

The analysis of expressivity data revealed a moderate baseline performance, with an intercept of 0.611 [0.52, 0.705]. Participants demonstrated a promising starting level of expressiveness. Over the intervention period, there was a slight improvement in expressivity, as evidenced by a time effect of 0.07 [-0.09, 0.24], indicating a small positive trend of 0.2% over time. At baseline, participants from Hannover performed slightly better (-0.11, [-0.33, 0.12]) with a 82% probability. The amount of time spent on homework had a small positive with a probability of 89% influence on expressivity improvement (0.1, [-0.06, 0.25]). At difficulty level 1, the participants from Hannover continued to have better scores on expressivity (-0.18, [-0.45, 0.09]). Furthermore, cognitive reserve had a postive impact on expressivity performance (0.12, [-0.02, 0.25]). Participants with higher cognitive reserve scores demonstrated better expressiveness. Similarly, higher perceptual abilities were associated with more expressivity (0.13, [-0.01, 0.26]). Additionally, higher general sophistication showed a positive effect on expressivity at the more challenging version of "Ode to Joy" (0.1, [-0.04, 0.24]), indicating that participants with greater general sophistication performed better in expressivity.

The findings suggest that participants made clear improvements in articulation, dynamics, pitch, and expressivity, while rhythm and fluency remained relatively stable. For and overview of possible predictors impacting piano performance at different levels of difficulty, see Table A.1. Only a few demographics could be confirmed to potentially predict the performance in various piano-related skills. Overall, it could be suggested that younger people with more cogntive reserve and a higher CogTel achieve higher scores in different aspects of piano performance. While people who describe themselves to have more emotions toward music, more singing and perceptual abilities, and a higher general musical sophistication reach higher scores when performance is tested, those who have less of all improve better over time. To generalize these observations the analysis is repeated on a potential *musicality* factor which could be extracted through factor analysis.

4.3 One General *Musicality* Factor

The factor analyses aimed to explore the underlying structure of musical abilities and identify patterns of shared variance among the individual musical variables across different time points and dicfficulty levels. The primary goal was to simplify the complexity of the data and extract a single latent factor, referred to as *musicality*, that represents participants' overall musical performance. Aditionally, the factor analyses investigated potential changes in musical abilities over time and under varying difficulty levels. Three seperate analyses were conducted: one at the baseline and the easy version of "Ode to Joy" ($FA_{0,0}$), another after 12 months of intervention and the easy version ($FA_{1,0}$), and the third after the intervention and the more challenging version of "Ode to Joy" ($FA_{1,1}$).

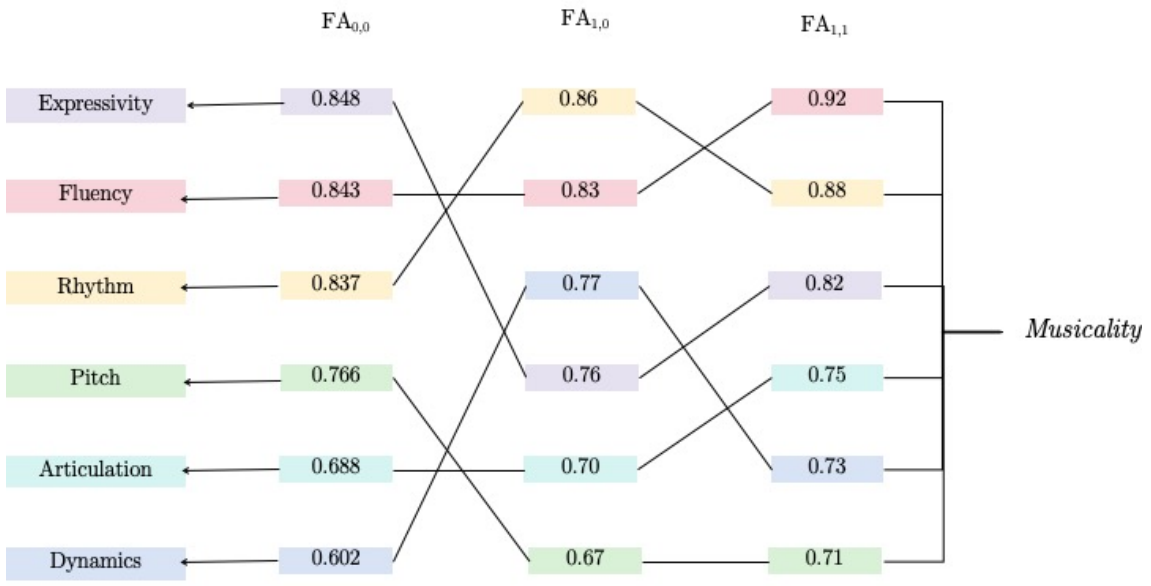


Figure 4.2: Factor Loadings on *Musicality*

Results from all three factor analyses consistently indicated the presence of only one dominant factor, which was labeled as *musicality*. This single latend factor accounted for 59% of the total variance in the data for both $FA_{0,0}$ and $FA_{1,0}$, and 65% of the total variance in $FA_{1,1}$. The ovserved musical variables (articulation, dynmaics, rhythm, fluency, pitch, expressivity) all shoed positive and moderate to strong factor loadings ranging from 0.602 to 0.848 in $FA_{0,0}$, from 0.67 to 0.86 in $FA_{1,0}$, and from 0.71 to 0.92 in $FA_{1,1}$ (see Figure 4.2). These factor loadings indicated a robust positive relationship between each musical variabel and the underlying *musicality* factor, suggesting that each variable contributed to participants' overall musical performance. Regarding changes in the musical variables over time, participants' *musicality* scores showed stability throughout the intervention and difficulty

levels. Notably, a considerable increase in the stability of one *muscality* factor was observed after the intervention with the challenging version of "Ode to Joy" ($FA_{1,1}$). The measures of fit, such as RMSEA index and Tucker Lewis Index of factoring reliability, indicated that all three factor analyses provided a goof fit to the data. The highest Tucker Lewis Index of factoring reliability was observed in $FA_{1,1}$.

After extracting the *musicality* factor, the factor loadings were multiplied with the original data and summed up to calculate a *musicality* score for each participant. These *musicality* scores ranged from -0.8 to 6.5, with higher values indicating a higher level of overall musical abilities and lower values suggesting a lower proficiency in multiple musical aspects. Subsequently, the *musicality* scores were analyed to investigate potential impacts of demographic factors. At the beginning of the intervention, the intercept for *musicality* was estimated at 2.9 with a 95% CI of [2.81, 2.99], representing the average *musicality* score for participants. A clear time effect was observed, indicating that participants' *musicality* scores decreased by approximately 0.37 units [-0.49, -0.24] from the beginning to after the intervention with the easy version of "Ode to Joy" (see Figure 4.3).

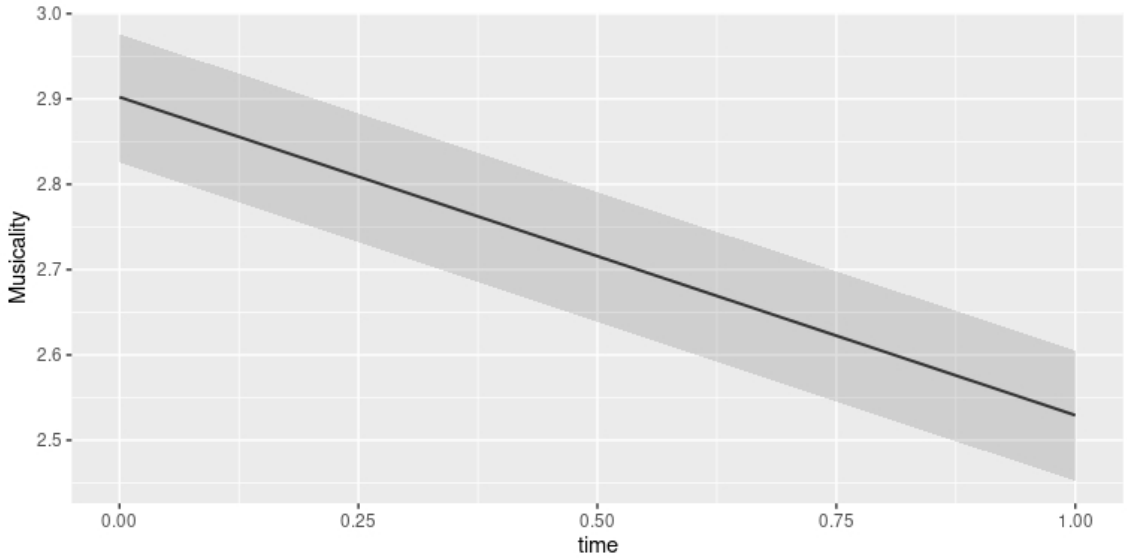


Figure 4.3: Predicted Values of *Musicality* over time

The results further revealed only a few demographic predictors that influenced participants' *musicality* scores. Initially, younger participants scored slightly higher in *musicality* (-0.08, [-0.18, 0.01]). Partcipants with higher cognitive reserve scores achieved better overall scores (0.10, [0.01, 0.19]). Aditionally, participants with higher CogTel scores obtained higher *musicality* scores (0.08, [0.00, 0.18]). The effect sizes are small but clear. Over time, participants with more cognitive reserve and higher CogTel scores showed less decrease in *musicality*

scores (cogntive reserve: 0.19, [0.06, 0.32]; CogTel: 0.19, [0.07, 0.32]). Similarly, the scores of younger (-0.22, [-0.35, -0.09]) and male (0.2, [0.09, 0.32]) participants diminished less. Participants with higher income and those who spent more time on their homework also experienced a smaller decrease in *musicality* scores (income: 0.16, [0.03, 0.29]; homework: 0.34, [0.21, 0.47]). Moreover, participants who were less engaged in musical activities and had fewer emotions towards music demonstrated a lower decrease of *musicality* scores after the intervention with the easy version of "Ode to Joy" (Musical Engagement: -0.25, [-0.38, -0.11]; Emotions: -0.37 [-0.49, -0.24]). Better singing abilities and more general sophistication in music also minimized the decrease of *musicality* over time (Singing abilities: -0.25, [-0.38, -0.12]; general sophistication: -0.18, [-0.31, -0.05]). In the more challenging version of "Ode to Joy", participants with higher education (0.2, [0.06, 0.34]), higher perceptual abilities (0.36, [0.22, 0.5]), higher CogTel scores (0.26, [0.11,0.4]) and more cognitive reserve (0.47, [0.33, 0.61]) performed better compared to other participants. Participants from Hannover (-0.31, [-0.45, -0.17]) and those who spent more time doing homework (0.23, [0.08, 0.28]) had a positive impact on musical performance. Aditionally, better singing abilities and greater general sophistication in music also positively influenced *musicality* scores after the intervention with the challenging version of "Ode to Joy" (0.19, [0.05, 0.34]; 0.26, [0.12, 0.41]).