Methods & Statistics
Research Paper
University College Utrecht
Fall 2002

Music & Memory

Sweta Hindocha
Laura van der Lubbe
Judit Koppany
Fiona Offerhaus
Floris van Vugt

Introduction

Carrying out a scientific research is always a challenge. One must combine attainability with novelty. Moreover, when working in a team each participant must show equal enthusiasm toward the topic, in order to attain the highest efficiency. Our research group also faced these two problems when trying to decide on a plausible research question for our investigation.
First, the idea of examining the relationship between the nature's of people's handshakes and their projected image came up. After careful consideration we turned down the idea, as we found almost impossible hard to conceptualize the nature of one's handshake. This was a clear crash between novelty and attainability. As we went on brainstorming, the proposition of investigating people's reaction to graphic images, that portray war situations, was made. Yet, even though some of us found it intriguing to see the correlation between previous experience (has the subject ever been exposed personally to war situations?) and current reaction (level of anxiety), others were not convinced that this can be measured objectively enough and awaited a more 'tangible' application. Thus, the quest for a new possibility began again. Finally, the idea of musical influence on intellectual performance was suggested. In order to overcome the time scarcity of the possible subjects, we decided to replace the several-hour-long intelligence test, by a memory examination. Hence, we narrowed the focus of our investigation from the interplay of music and intellectual performance to the one of music and short-term memory.

Operationalization

In order to test the effect of music on short-term memory, we decided to use a pretest (no music) and posttest (with music) system that would clearly show the influence of introducing music as an environmental factor. As we hoped to test the influence different types of music have, we chose two distinct genres: classical and rock. Our experimental setting looked the following:

Pretest Classical	Pretest Rock
↓	↓
Posttest Classical	Posttest Rock

To eliminate the contamination of preference (some people might be more used to classical; respectively, rock music), the participants were asked to indicate their liking at the end of the test. Hence, there were four distinct subgroups, where the first word indicates the test group and the second shows the preference:

Posttest Classical		Posttest Rock	
↓	↓	↓	↓
Class(Rock)	Class(Class)	Rock(Rock)	Rock(Class)

Testing the short-term memory took place in the form of memorizing 20-20 Hungarian words, using the raw scores (number of correct answers). We chose Hungarian, because it is a language dissimilar to other languages. Hence, no one would have an advantage (by using his knowledge of other languages) when learning the words. After setting the founding stones of our operationalization, we entered the next phase of our experiment: the specification of the research question.

Research Question:

The goal in our research is to determine whether, and how different types of music (rock and classical) influence short-term memory. In other words, we are interested to see whether listening to classical music (corresponding to the general belief) enhances performance and listening to rock music worsens it.

Hypothesis:

We expect that our data will confirm the general standpoint. Namely, subjects who listen to classical music will perform better than the people in the rock group.

Our nul- hypotheses is Ho: μ pretest-posttest $_{classical}$ - pretest-posttest $_{rock} = 0$
The alternative hypothesis is Ha: μ pretest-posttest $_{classical}$ -pretest-posttest $_{rock} \neq 0$
We decided to use the mean for the difference between pretest-posttest, since this variable includes the paired data.

Furthermore, we anticipate that the above-described hypothesis will be affected by the music preference of the subjects. Accordingly, we expect the difference between pretest-posttest to differ with preference. Subjects who have to do the test with the music (rock or classical) will have a lower difference than people who are not.

Ho: The μ of subjects in either rock(rock) or class(class) – either rock(class) or class(rock) = 0
Ha: μ of subjects in either rock(rock) or class(class) – either rock(class) or class(rock) $\neq 0$

Methods

In order to investigate the hypotheses a research experiment was set up that would yield the highest internal validity possible. It was constructed in such a way that the obtained results were minimally affected by any extraneous factors. Some extraneous effects, however, were either impossible to rule out or it would cause a loss of other data if altered. However, many things were taken in consideration and after weighing all the advantages and disadvantages we choose the following set up.

Procedure

The way the experiment was conducted is as follows. Two tests were used in the experiment. Each test makes use of a sheet of paper with 20 Hungarian words written on it with the English translation written behind each Hungarian word[1]. The reason why the Hungarian language was chosen was because this language is very unlike any other languages. It has no similarities with English or with other commonly spoken languages. Therefore, it was the language that would most likely be equally difficult for each subject participating. The words that were chosen were basic English words, such as dog, school etc. These were chosen to avoid that native English speakers would have an advantage, since it might be difficult to remember long, uncommon English words for a non-native speakers. Each subject was given four minutes to try to learn all the translations of the Hungarian words by heart. After these four minutes, the sheets were collected and new sheets were given. This second sheet has the same 20 words written on it only in a different order and without English translation. The subjects were now given 2 minutes to try to fill in the English translation behind each word. The time slots were determined after a test-of-the-test was held. This test, which was conducted by the experimenters, showed that these time slots were just enough to keep the participants interested. It showed that, within four minutes, the participants could learn all the words but the time was so short that it was positive that these words would be restored in the short-term memory of the students. Furthermore, it was clear that two minutes were enough to fill all the words in and think for a short while. This first test was made by a randomly drawn sample of students, consisting of 40 students. All of the students were people studying at University College. Ages were not asked, since these are not important to our research, however it can be assumed that the students' ages were in the range from 18-22. The way the sample was obtained was by creating a list of random numbers using a graphic calculator. A student magazine from University College named 'The Boomerang' was used to function as device from which the random numbers would be matched with a student. This student was approached and asked whether he or she wanted to participate in the experiment. E-mails were sent out a day before the experiment to remind everybody to come and to give participants the opportunity to cancel. Eight people cancelled [2] and substitutes were found using the same selection method as used drawing the original sample. The tests were conducted in the following way. On a Wednesday afternoon at 16:15 two classrooms in Voltaire were used to divide the groups.

[1] See appendix 1

[2] Various reasons were given for canceling. However, taking these reasons and the goal of our research into account it was concluded that this needed no further consideration in the course of the research. No pattern or reoccurring reason was found.

Voltaire is a building on University College. The rooms are square shaped with large windows at one side. The temperature in the building can be considered normal. Every student was given a number and again using the graphic calculator a list of random numbers was obtained. Using these numbers the group was divided in two groups that were theoretically equal, due to the random sampling.

Condition and specifics

The groups were told that the two persons who performed best on both test (so the two highest means) would be awarded with a movie coupon. This would presumably enhance the quality of the data, since people will do their best to perform as well as possible. Also, a movie coupon would be given to a random person. This approach was chosen to make sure that people who did not perform too well on the first test were not discouraged by this and would still put effort in making the second test. Each group, after the division now consisting of 20 students, was given this first test. Both the learning as the filling in of the translations was done in silence. That is, of course, as silent as possible because it cannot be prevented that the participants hear some sounds. However, since the two classrooms were exactly next to each other, we assumed that the circumstances concerning noise were similar. Other circumstances that might cause a difference in performance in the two groups were also equal, such as for example the temperature in the room. Since the tests were conducted at the same day, at the same time, it was also possible to rule out tiredness as a factor that caused a difference between the groups. After this first test, papers were collected and the students were explained what was to be expected in the second test. Although it was assumed that it was likely that people would perform less on a second test even if the test was identical to the first one (since people would be more tired and less enthusiastic than in the first test), it was decided that the best approach was to do the tests directly after each other. Since, the weaker performance due to tiredness was partially counteracted by a better performance due to the better familiarity that the subjects now had with the testing system. Also, it was thought that, dividing the tests in two tests on two separate days would yield a lower number of participants who were willing to cooperate. It was furthermore realized that it might be the case that people who did show up the first time would not show up the second, which would cause a loss of valuable data. Furthermore, even though an effect of performance was expected due to the fact that the two tests immediately followed each other, this effect was expected identical for both groups. Therefore, any significant evidence that might be found would have to have a different cause. The second test was exactly the same as the first test, only different words were used. So, for both groups this second test was handed out which did not differ between the two groups. The situation in which they made the test, however, did differ this time. The first group had to learn the words, as well as fill them in, while continuously listening to classical music. The second group had to learn as well as fill in the words while continuously listening to rock music. The definition of rock music in this experiment was non-vocal, up-beat music with much base in it. The definition of classical was also non-vocal, slow and without a beat or a base. The reason why non-vocal music samples were chosen was to prevent that the voices distract people too much. Furthermore, on the forehand it was checked that the volume level was equal for both groups. The loudspeakers were placed on the table in the middle of the room. The students participating sat in approximately a circle around it. It is highly

unlikely that one student perceived the level of music to be much higher than a student sitting in another place. In addition, due to the way the students were seated, it was not possible for them to look at the paper of a person next to him. The conductors of the experiment were present at all times, which also made it impossible to cheat.
In conclusion, many things were taken into consideration when the experiment set up was to be determined.

Explanation of terms

In the data analysis we used different short terms. We will explain what these terms mean.

- Pretest: the first test, which was identical for all 40 subjects, namely the same words and no music.

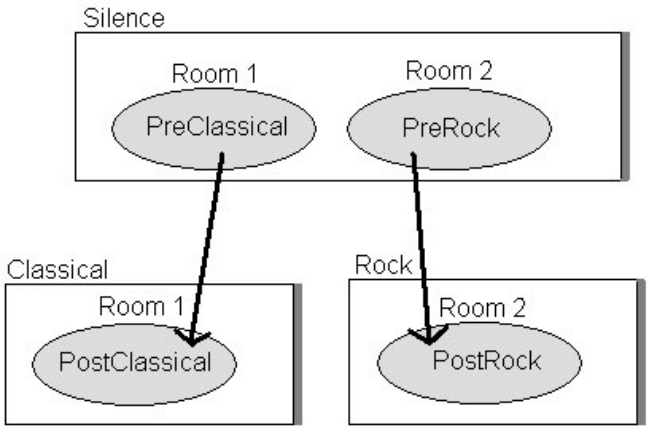
- Posttest: the second test, in which the group was divided into two, where one group had to listen to classical music and the other group had to listen to rock music. The words were the same for both groups.

- Rock: the subjects which were assigned to listen to rock music.

- Classical: subjects which were assigned to listen to classical music.

- Rock(Rock), Rock(Classical), Classical(Classical), Classical(Rock):
The assignment of the subjects to which group along with their preferences. The first word identifies to which group the subjects were assigned, the second word, the word between brackets, shows us their preference. For instance, Rock(Rock) means that the subjects were assigned to listening rock music and that they also prefer listening to rock music.

Test-of-the-test: trying the test ourselves in order to see how much time the subjects need to learn the words and to make the test.



Experiment Results

Group Merging

Due to the problems we experienced and explained before, we had to merge the subjects from the first Rock test and the subjects from the second Rock test (a week later). Although we did our best to keep all circumstances exactly equal, we decided that we required some tests to ensure that there are no significant differences between the two.

The scores for each test consist of the number of correct answers.

We performed a two-sided two-group T-test for both the Pretest and the Posttest, comparing the scores of both groups. The results are displayed below:

Table 1: Group Statistics

	Test Group	N	Mean	Std. Deviation	Std. Error Mean
Pretest Score (no. correct)	Rock1	10	14.10	5.021	1.588
	Rock2	10	18.10	2.283	.722
Posttest Score (no. correct)	Rock1	10	14.50	4.950	1.565
	Rock2	10	16.40	2.875	.909

Table 2: Independent Samples T-Test

	Levene's Test for Equality of Variances		t-test for Equality of Means		
	F	Sig.	t	df	Sig. (2-tailed)
Pretest Score (no. correct)	10.140	.005	-2.293	12.568	.040
Posttest Score (no. correct)	1.169	.294	-1.050	14.453	.311
Posttest Score (no. correct)	1.169	.294	-1.050	14.453	.311

As one can see, there is a significant difference between the pre-test scores. However, this is not all there is to it. Consider the following test, comparing the average difference between the pre- and posttest scores for both groups:

Table 3: Group Statistics

	Test Group (Classical /Rock)	N	Mean	Std. Deviation	Std. Error Mean
Difference between Pre- & Posttest	Rock1	10	-.40	2.366	.748
	Rock2	10	1.70	3.592	1.136

Table 4: T-Test

		Levene's Test for Equality of Variances		t-test for Equality of Means				
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
Difference between Pre- & Posttest	Equal variances assumed	.846	.370	-1.544	18	.140	-2.10	1.360

This shows a difference that is not significant.
On the basis of these tests, we decided we were justified to use the data from the week later, considering:

1) We had done everything we could to keep the circumstances identical in both experiments.
2) We had not observed any differences in environment or the subject's responses, except for the more detailed instruction.
3) The second variable, the calculated difference between the pre- and the post-test constitutes:
    a. the data we want to use in our analyses, as we are comparing the pre- and the post-test, therefore the individual values of the pre- and the post-test are less important, and
    b. a more reliable measurement, because it is based on two measurements within one subject (a paired test).

General Statistics

Now we can present the results of the experiment after we merged the groups and left out the wrong samples as described before.
The set of data we have now is used to calculate several new variables for each subject. Please refer to the appendix for an overview of the variables.
One variable however is of especial importance to our enterprise, and that is TestDiff, which is the difference between the Pretest and the Posttest.
Formula: (Testdiff) = (Pre) – (Post).
Note that the TestDiff variable is positive if the Posttest is lower than the Pretest. So it is the decrease of the test performance between the Pre- and the post-test.

Exploration

Table 5: Descriptives of the final population

	Pre	Post	TestDiff
Mean	16.08	14.73	1.35
Variance	13.507	17.230	10.079
Std. Deviation	3.675	4.151	3.175
Minimum	5	3	-5
Maximum	20	20	9

Table 6: Case Summaries

		Pretest Score	Posttest Score
Rock(Rock)	N	15	15
	Mean	16.73	16.00
	Std. Deviation	3.807	2.928
Class(Rock)	N	13	13
	Mean	15.00	13.54
	Std. Deviation	3.215	4.274
Rock(Class)	N	5	5
	Mean	14.20	13.80
	Std. Deviation	5.630	6.611
Class(Class)	N	7	7
	Mean	18.00	14.86
	Std. Deviation	1.155	4.298
Total	N	40	40
	Mean	16.08	14.73
	Std. Deviation	3.675	4.151

Figure 1: PreTest Score

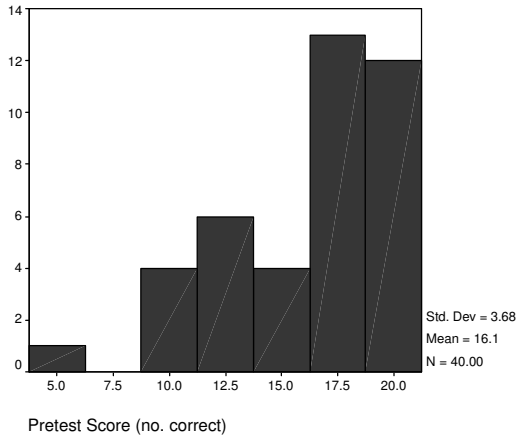
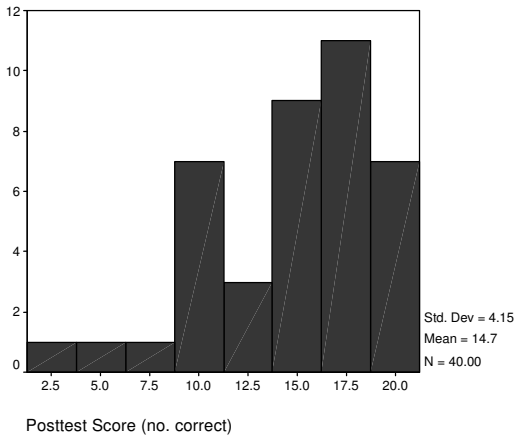
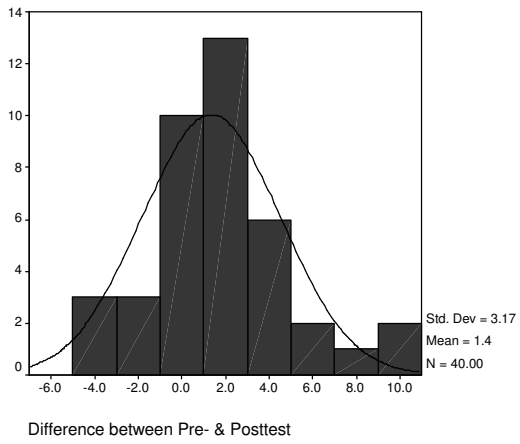


Figure 2: PostTest Score



The two graphs above indicate that we have found a distribution that seems more skewed than a normal distribution. However, one must remember that the maximum score in this test in 20, and a considerable number of people came at least close to that number. Therefore, perhaps the distribution of scores is normal in reality and people would have been able to learn more words than the 20 we gave.
The graph below shows that it is quite plausible that the variable TestDiff is approximately normally distributed.

Figure 3: TestDiff. Score

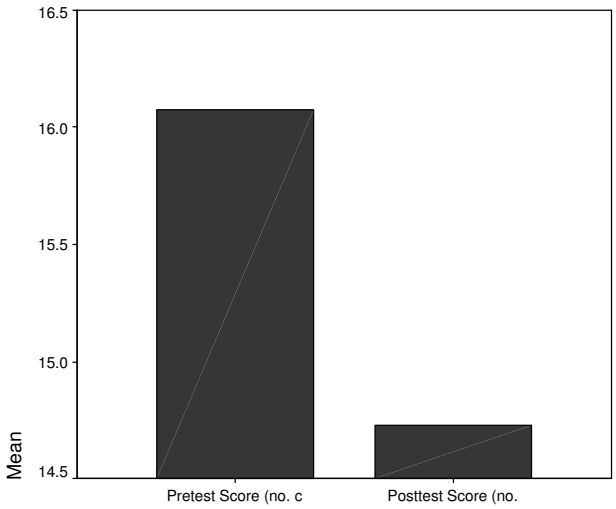


The table depicted below shows that there is a difference in mean scores for the pre- and the post-test. However, several environmental conditions were different between the pre- and the post-test:

a) whether there is music or not.
b) the people had done the test already once, so they knew how it would go.
c) the tests were different, and many people observed the second test to be more difficult than the first one.

Therefore, even if there would be a numerically significant difference in scores between the pre- and the post-test, we would not be allowed to draw conclusions from that with respect to our research enterprise.

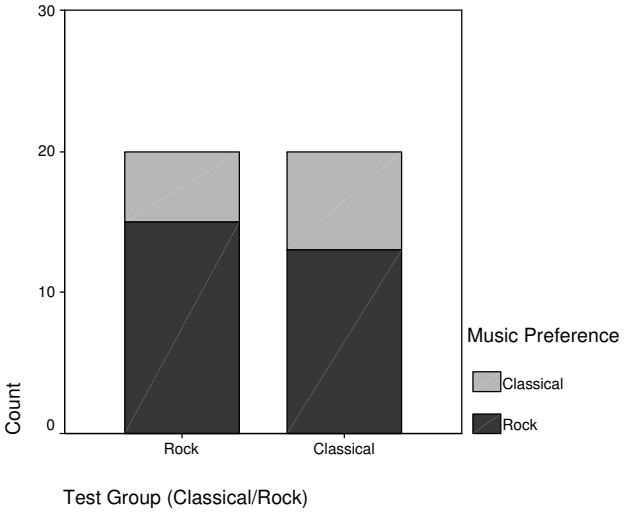
Figure 4: PreTest and PostTest average scores



Some Peculiarities

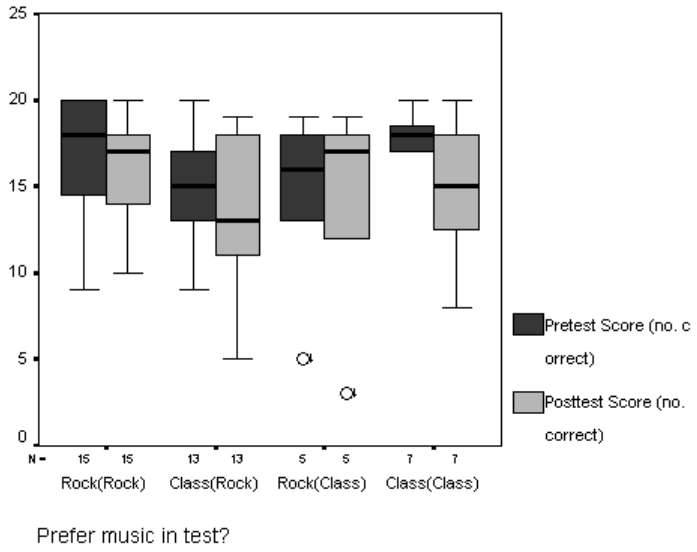
The graph below reveals that there is a higher amount of people who prefer classical music in the Classical group, in spite of our random assignment. Therefore, perhaps there is an influence of the kind of music played during the test has an influence on the kind of music they say to prefer.

Figure 5: Music Preference per Group



The next graph shows the distribution of both the scores with respect to the new variable at_place2, which is a combination of the group the people are in and their musical preference.
Rock(Class) means that one is put in the Rock group, while his or her preference is actually classical music.

Figure 6: Boxplots per Group and Preference



First of all it is apparent that there is quite a lot of variation in the pretest, although all the subjects are in exactly the same conditions there. This variation is perhaps even larger than the variation in the posttest scores. This would lower our chances of being able to make a justified conclusion from the data given. We will come back to this problem later.
Second of all we see that the posttest scores are lower than the pretest scores.
a. this is in accordance with what people told us; they reported that in their experience the second test (which was the same for all of them) was harder than the first test.

Relevant Results

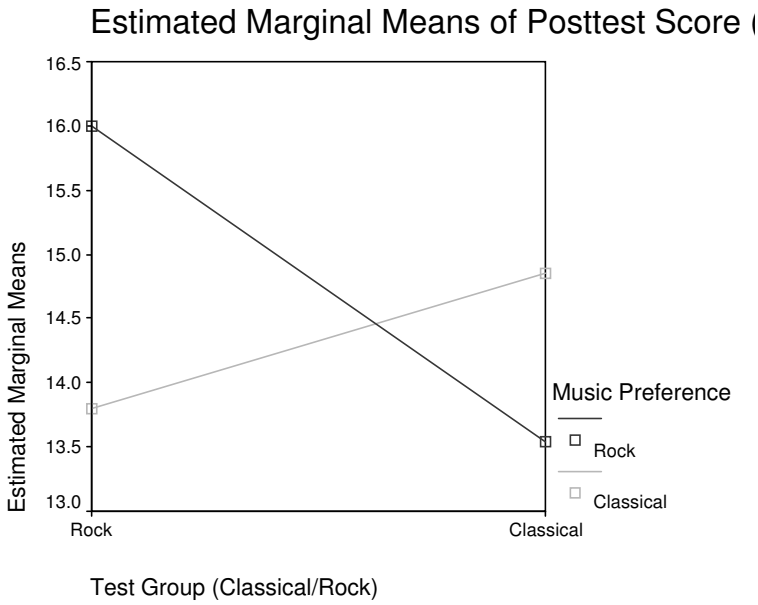
First of all, we performed an ANOVA analysis on the interaction between the group the subjects were in and their musical preference, and how that affected their PostTest score:

Interaction on PostTest

Table 7: ANOVA PostTest Group and Preference

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	47.087	3	15.696	.904	.449
GROUP * PREF	25.452	1	25.452	1.466	**.234**
Error	624.888	36	17.358		

Figure 7: Interaction Plots of PostTest Score per Preference and Group



As one can see, even though the interaction seems to be exactly what we expected, the interaction is not significant (P-value = .234).
The same analysis we perform on the TestDiff variable:

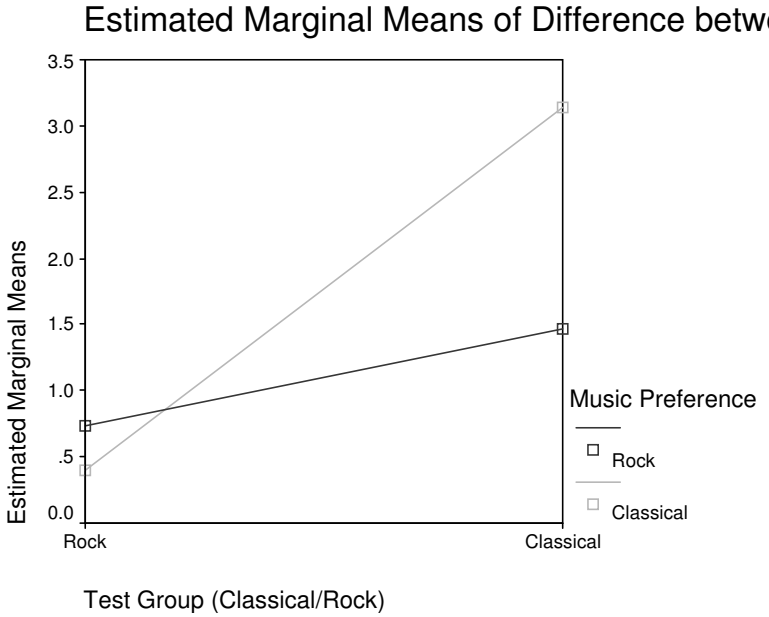
Interaction on TestDiff

Table 8: ANOVA TestDiff Group and Preference

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	32.879	3	10.960	1.095	.364
Intercept	67.678	1	67.678	6.764	.013
GROUP	24.768	1	24.768	2.475	.124
PREF	3.735	1	3.735	.373	.545
GROUP * PREF	8.344	1	8.344	.834	**.367**
Error	360.221	36	10.006		
Total	466.000	40			
Corrected Total	393.100	39			

a R Squared = .084 (Adjusted R Squared = .007)

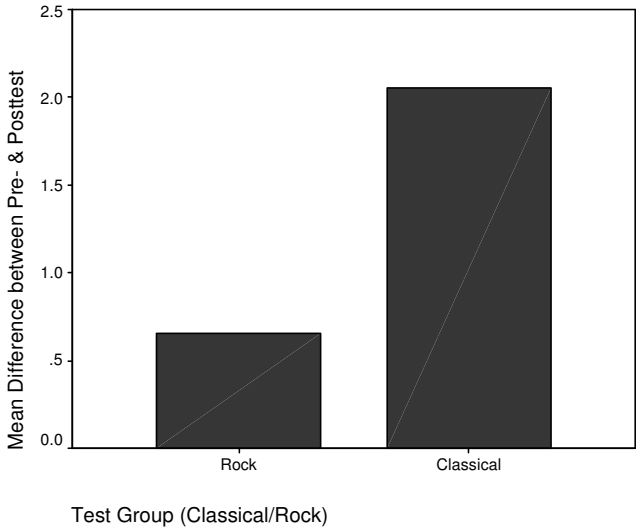
Figure 8: Interaction Plot of Testdiff per Group and Preference



These figures show that the interaction as plotted looks precisely the way we expected it to be. But the ANOVA analysis showed that both interactions are not significant (neither would be using an already not very severe significant level like 10%).

Therefore, in the final analysis, we will have to ignore the musical preference.
In this analysis we will use the variable TestDiff as much as possible, since this variable is stronger than the pre- and post-test averages separately, since it makes use of the fact that we have paired data (every subject participated in both the pre- and the post-test).

Figure 9: Average TestDiff per Group



Although the graph below seems to indicate that in the Rock group the average decrease in scores between the Pre- and the Post-test is smaller. It would seem that Rock is therefore of a better influence than Classical music. However, as the tables below show, this difference is not significant either on a 10% level.

Table 9: Group Summaries

	Test (Classical/Rock)	Group N	Mean	Std. Deviation	Std. Error Mean
Difference between Pre- & Posttest	Rock	20	.65	3.150	.704
	Classical	20	2.05	3.120	.698

Table 10: T-Test PreTest and PostTest

	t-test for Equality of Means						
	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
						Lower	Upper
Equal variances assumed	-1.412	38	.166	-1.40	.991	-3.407	.607
Equal variances not	-1.412	37.996	.166	-1.40	.991	-3.407	.607

assumed |

Conclusions

The goal in our research was to determine whether, and how different types of music influence short-term memory. From our results can be seen that music indeed has an influence on short-term memory, since the pre-tests scores in both groups are higher than post-test scores. The pre-test score is namely 16,88 and people scored only 14,73 at the post-test. However, this could also be caused by other factors, such as the difficulty of the post-test or loss of concentration due to tiredness. There is also much more variance in the post-test scores. The reason for this could be that some people are used to study with music and some are not. This might be a question to ask when the research is done again. We predicted that students could concentrate better when they listen to classical music. However, when we look at the graph of the mean differences between pretest and posttest, it can be seen that the difference for classical music is much higher. This means that people who listened to classical music scored much lower on the posttest compared with their pretest than people who listened to rock music. This of course is against our assumptions. Although rock music does seem to have a better influence on the scores than classical music, this is not significant. The two-tailed T-test gives us a significance of 0.166. This is therefore not significant.

Our hypothesis was:
Our nul- hypotheses is Ho: μ pretest-posttest classical - pretest-posttest rock = 0
The alternative hypothesis is Ha: μ pretest-posttest classical - pretest-posttest rock ≠0
We decided to use the mean for the difference between pretest-posttest, since this variable includes the paired data.
The mean for the difference between pretest-posttest classical is 2.05 and the mean for the difference between pretest-posttest rock is 0.65. The two-tailed T-test for the means has a very high significance of 0.166, which leads to the conclusion that we cannot reject our null-hypothesis.

Another hypothesis was:
Ho: μ of subjects in either rock(rock) or class(class) - either rock(class) or class(rock) = 0
Ha: μ of subjects in either rock(rock) or class(class) - either rock(class) or class(rock) ≠ 0
As can be seen from our analysis in SPSS is that people who are put in either the classical or rock group and have the same preference as the group they were put into, score higher on the post-test than people who are in the ' wrong' group. Thus, for example people who are in rock, and prefer rock have higher scores than people who are also in rock, but prefer classical music. The mean score of the Rock(Rock) group is 16, whereas the mean score of the Rock(Class) group 13,80 is. The same goes for classical music. However, the people who were put in the class with the music they preferred also performed better on the pre-test. This of course is rather odd, since the circumstances in the pre-test are the same for everyone. Since they already scored higher on the pre-test, is it not strange that they also scored higher on the post-test. With this supposition we have to reject our assumption that preference has an influence on short-term memory.

Besides, there are big outliers, namely a score of five and three, in the rock(class) group. This can be seen very well in the box plots of all the different groups. These quite extreme outliers have a big influence on the mean of this group. It is possible to do an analysis ignoring these outliers, but it seems difficult to come up with a reason for deleting these outliers. One reason could be that these people have dyslexia, but it is difficult for us to find this out. Or they just might be very bad in learning words.
Another thing that has to be mentioned is that preferences might have influenced by the music people just heard, while we explicitly told them too make a decision regardless of the music they just heard.
Although there does seem to be an influence of music preference and test scores, as can be seen in the graph of the estimated marginal means, this influence does not seem to be significant. It namely has a significance of 0.234, which is rather high. And the estimated marginal means of difference between subjects is not significant too, namely 0.367. You would need an extreme significance level to make that significant. So again we cannot reject our Ho-hypothesis.

Discussion and Evaluation of Experiment

While carrying out our experiment, several problems came to surface that had not been addressed beforehand that clearly demonstrated the conflict between the theoretical approach of the design and the practical manner of the investigation, itself. The three major problems included the lack of pre-defined and written instructions, the consistency in difficulty of the pre- and posttest and the maturation effect.

Firstly, during the first set of our experiment, there were no pre-defined and printed instructions that the supervisors read aloud to the subjects. In the classical group, it was stressed the subjects were required to memorize the English translations and they were not supposed to learn the spelling of the Hungarian words. In the rock group, this was not mentioned. Consequently, there were several subjects in the rock group, who studied both the English and the Hungarian words resulting in a significantly lower performance. Not only did they worse on the pretest, but also (as predictable) by the posttest they had a clear understanding of the procedure, resulting in high accomplishment. Thus, the score difference between the pre-test and the posttest was remarkably lower in some cases in the classical group, than in the rock. In order to tackle this problem, the experiment was carried out a second time the following week (same day and same time slot). New subjects were randomly sampled, and they repeated the experiment with rock music. This time, the supervisors were asked to pay attention when defining the task and to use the exact same wording as one week ago in the classical group. The results gathered in the second set (rock 2) were used as replacements for the ones who clearly lacked the understanding of the procedure in the first set (rock 1)-those subjects verbally indicated their confusion after having ended the experiment.

Many other participants voiced a second remark, the increased difficulty of the posttest when compared to the pretest. They argued that the second set had words with more syllables; furthermore, distinguishing between the words, virag (flower) and vilag (world), gave them a hard time. However, this increase is difficulty is not a confounding variable, as the focus of our experiment is not the examination of the difference between the no music and music environment, but the investigation of the effect different kinds of music have. As all of the subjects both in the rock and the classical group were given the same pretest and posttest, both of their performances (rock versus classical group) are compared on the basis of two environments with only *one* differing variable that is the type of music.

Finally, even if the subjects were clearly asked to memorize the Hungarian words only, a maturation effect cannot be excluded, as by their test-taking skills must have naturally improved from the retest to the posttest. However, when following the same line of reasoning as in the case of difficulty consistency, it can be stated that the influence of this improvement was present with each subject. Accordingly, it does not contaminate our research, and the results can be viewed as highly reliable.

In short, some problems were present during the actualization of our investigation; nevertheless, they do not lower the worthiness of the experiment. As they were either resolved by changing the conditions, or they were of not confounding nature in general.

Generalization

The purpose of conducting a research is often to generalize the results to a larger population. That can be any group: woman, men, teenagers or alcoholics. In our research we wanted to generalize the results to university students.
In order to be able to generalize, the external validity needs to be high. A minimum prerequisite to achieve high external validity is testing plenty of subjects. The amount of sufficient subjects differs per experiment. We chose to test 40 subjects. We chose this amount because we wanted sufficient subjects, but we also had to take into account the amount of time the test would occupy. That is why we thought 40 subjects would be sufficient.
However, our test results did not prove us right. As told before, we did not find any sufficient correlation between the results. There was some correlation, but the level was not high enough. This could be because of the insufficient number of subjects, small differences may seem bigger than they in fact are.
However, it could of course also be that there simply is not an effect of music on memory. If that is the case, an infinite number of subjects would not have shown a significant effect either.
We still feel that a larger number of subjects would give up more clarity whether there is an effect of music or not.
In short, as our external validity does not seem to be very high, we conclude that we would need more subjects.

Suggestions for Improvement

As mentioned before, we came across some unforeseeable problems. Some of them we were able to mend. For instance, we found out that in both groups we did not give the same instructions. The result was that one group did not know that they only had to learn the English words, and not the Hungarian words. This spoiled some of our data. As soon as we found out which data was not in accordance to what we were aiming at, we organized a new experiment. We replaced the 'wrong' data collected from the first experiment with the data from the second experiment. We know now that next time we should write down the instructions for ourselves, to make sure that they are clear and the same. We thought that it would be better if we would read the instructions out loud, considering that if we would hand them out to the subjects not all of them might read it.
Another aspect which needs improvement is the amount of subjects we tested. 40 Subjects appeared not to be sufficient, considering we were not able to draw sincere conclusions. The level of correlation was not high enough. The suggestion for improvement for the next time is to ask more subjects, up to an amount where the effect of outliers can be demolished.
Another problem we came across was that our subjects came up to us mentioning that the second test, the test with the music, appeared to be more difficult than the first one. They mentioned that the words were more difficult. The exact reason why it was found more difficult, except for the fact that there was music playing of course, has already been explained earlier. However, considering everyone got the same tests, the effect of that inconvenience is not that severe. We would be worse off if we would have two different tests for the second round, because then the reliability would go down.
We could have taken this effect away if we would have done a better test of the test itself. We did test the test on ourselves, but then we only focused on the amount of time we would need to learn the words. We forgot to look at the difficulty of the words. So that is another thing we could improve next time.
So apparently we have come across several problems during our research, some bigger than others. We have tried to rectify them as much as possible. However, these mistakes just show how realistic this research in fact is.

A Final Word

What has become very clear to us is that research apparently cannot be predicted. Neither the outcome of the results, nor the course of the experiment is predictable.
The results we had expected did not come out, for instance the calming effect of classical music did not seem that big, it appeared that the rock group even performed better. We have learned a lot from this.
We also learned a lot from the mistakes we made when executing the experiment. We now see flaws in our organisation while normally we would not even have noticed. However, our results pointed them out to us. Now we have learned from our mistakes.
We do feel we have to emphasize the fact that we enjoyed doing this research. Especially doing the test itself, which is of course where we put all our effort in, was very much fun to conduct. The most significant was how well students were willing to participate, and how well they studied. It was also very nice that they took the time to give us feedback about the tests. Without that we would probably have left in some errors, luckily we were able to take some of them out.
Another thing we liked about conducting this experiment was our subject. The influence of music was very interesting, considering we belong to the same target group of our research and so the results would also have some significance for us too.
In short, we enjoyed conducting this research despite that we made some mistakes. We can only learn from them.

Appendix

- Experiment set-up
- List of words
- List of word-explanation
- The Resulting Data

Methods Experiment Set-Up

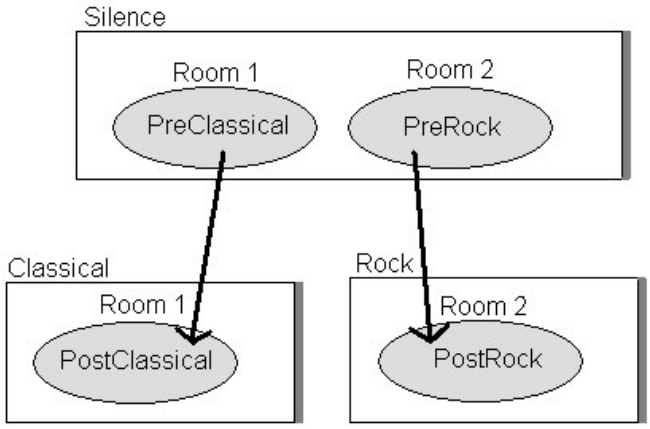
Wednesday November 20th, 2002
4.15 pm
Voltaire

1. Subjects are randomly assigned to one of the two groups. This is achieved by letting them take (without watching) a small paper from a box which contains 20 small papers in one colour and 20 of another colour.
2. Each group has its own classroom.

   Experiment (two of the experimenters will have to be present in each classroom)
   a. The experiment is explained to the subjects.
   b. Each subject is given a sheet of paper with 20 Hungarian words, with the written side downwards. He is not allowed to turn over the paper until a mark is given.
   c. 4 minutes are allowed to learn as many words as possible.
   d. After 4 minutes, a signal is given and the subjects must turn their paper over.
   e. A test will be handed to them while the papers with the Hungarian words are removed from the tables. They are not allowed to turn over the test until a mark is given.
   f. 2 minutes are allowed to fill out the test.
   g. After 2 minutes, a signal is given and the subjects must turn their paper over.
   h. The test will be collected.

3. Each group takes the experiment once without music and once with music playing in the background (rock for one of the groups, classical for the other).
4. Each subject is asked is preference (classical or rock) regardless of the music that was played just now when collecting the Posttest test.

Systematic Naming:



The sheets:

Words= The sheets with the Hungarian words and their English translation.
Test= The sheet with only the Hungarian words in a different order from the Words-sheet and an empty space for the subject's name.
Check=The words in test order
PreWords

List of Words

Pretest Words

1	**FA**	Tree
	HÁZ	House
	KUTYA	Dog
	CICA	Cat
5	**ANYA**	Mother
	APA	Father
	NŐVÉR	Sister
	BÁTY	Brother
	ASZTAL	Table
10	**SZÉK**	Chair
	ISKOLA	School
	VONAT	Train
	KÖNYV	Book
	REPÜLŐ	Plane
15	**KOCSI**	Car
	BICIKLI	Bike
	AJTÓ	Door
	HÍD	Bridge
	TOLL	Pen
20	**CIPŐ**	Shoe

PreTest Check Sheet

Name

Noname

1	**HÍD**	Bridge	
	KÖNYV	Book	
	BÁTY	Brother	
	CICA	Cat	
5	**SZÉK**	Chair	
	KUTYA	Dog	
	FA	Tree	
	HÁZ	House	
	ANYA	Mother	
10	**TOLL**	Pen	
	ISKOLA	School	
	NŐVÉR	Sister	
	VONAT	Train	
	BICIKLI	Bike	
15	**KOCSI**	Car	
	AJTÓ	Door	
	REPÜLŐ	Plane	
	CIPŐ	Shoe	
	ASZTAL	Table	
20	**APA**	Father	

PreTest

Name _____

1	**HÍD**	
	KÖNYV	
	BÁTY	
	CICA	
5	**SZÉK**	
	KUTYA	
	FA	
	HÁZ	
	ANYA	
10	**TOLL**	
	ISKOLA	
	NŐVÉR	
	VONAT	
	BICIKLI	
15	**KOCSI**	
	AJTÓ	
	REPÜLŐ	
	CIPŐ	
	ASZTAL	
20	**APA**	

Posttest Words

1	**VIRÁG**	Flower
	SZÁMÍTÓGÉP	Computer
	CSILLAG	Star
	HOLD	**Moon**
5	**NAP**	Sun
	VILÁG	World
	ÓRA	Watch
	NÉZ	To see
	SÉTÁL	To walk
10	**BIRTOKOL**	To have
	VÍZ	Water
	ABLAK	Window
	ZOKNI	Sock
	IDÕJÁRÁS	Weather
15	**RUHA**	Clothes
	VAN	To be
	POHÁR	Glass
	JÁTSZIK	To play
	PIROS	Red
20	**KÉK**	Blue

PostTest Check Sheet

Name

Noname

1	**VILÁG**	World	
	RUHA	Clothes	
	POHÁR	Glass	
	HOLD	**Moon**	
5	**PIROS**	Red	
	CSILLAG	Star	
	NAP	Sun	
	JÁTSZIK	To play	
	SÉTÁL	To walk	
10	**ABLAK**	Window	
	BIRTOKOL	To have	
	ZOKNI	Sock	
	ÓRA	Watch	
	SZÁMÍTÓGÉP	Computer	
15	**KÉK**	Blue	
	IDÕJÁRÁS	Weather	
	VIRÁG	Flower	
	VÍZ	Water	
	NÉZ	To see	
20	**VAN**	To be	

PostTest

Name _____

1	**VILÁG**	
	RUHA	
	POHÁR	
	HOLD	
5	**PIROS**	
	CSILLAG	
	NAP	
	JÁTSZIK	
	SÉTÁL	
10	**ABLAK**	
	BIRTOKOL	
	ZOKNI	
	ÓRA	
	SZÁMÍTÓGÉP	
15	**KÉK**	
	IDÕJÁRÁS	
	VIRÁG	
	VÍZ	
	NÉZ	
20	**VAN**	

List of word-explanation

Variable	Meaning
ID	A unique number assigned to each subject.
Pref	Which music the subject likes to hear best: 1=rock 2=classical
Study	Whether the subject usually studies with music: 0=no 1=yes,rock 2=yes,classical 3=yes,but other.
Group	The group to which the student was assigned: 1=rock 2=classical
Pre	The Pretest score in number of correct answers (out of 20)
Post	The Posttest score in number of correct answers (out of 20)
TestDiff	The difference between the Pretest and the Posttest. Formula: (Testdiff) = (Pre) – (Post).
Avg_Scor	The average of the Pre- and the Posttest score Formula: (Avg_Scor) = ((Pre)+(Post)/2)
Atplace	Whether the subject is hearing the music he likes to hear most (in general).
Atplace2	Whether the subject is hearing the music he likes to hear most and which group he is in. Formula: (Atplace2) = ((Pref*2)-1)+(Group-1).

The Resulting Data

ID	PREF	STUDY	GROUP	PRE	POST	TESTDIFF	AVG_SCOR	ATPLACE	ATPLACE2
aa	1	#####	1	17	16	1	16.5	1	1
ab	1	#####	1	19	19	0	19.0	1	1
ac	1	#####	1	12	14	-2	13.0	1	1
ad	2	#####	1	5	3	2	4.0	0	3
ae	1	#####	1	12	10	2	11.0	1	1
af	1	#####	1	11	14	-3	12.5	1	1
ag	1	#####	1	20	18	2	19.0	1	1
ah	1	#####	1	9	14	-5	11.5	1	1
ai	1	#####	1	17	18	-1	17.5	1	1
aj	2	#####	1	19	19	0	19.0	0	3
ak	1	0	1	20	18	2	19.0	1	1
al	1	2	1	19	18	1	18.5	1	1
am	1	2	1	20	20	0	20.0	1	1
an	1	3	1	20	15	5	17.5	1	1
ao	1	1	1	20	11	9	15.5	1	1
ap	1	3	1	18	17	1	17.5	1	1
aq	2	0	1	16	12	4	14.0	0	3
ar	2	3	1	18	18	0	18.0	0	3
as	1	1	1	17	18	-1	17.5	1	1
at	2	0	1	13	17	-4	15.0	0	3
ba	1	#####	2	14	12	2	13.0	0	2
bb	2	#####	2	17	17	0	17.0	1	4
bc	2	#####	2	18	14	4	16.0	1	4
bd	2	#####	2	17	8	9	12.5	1	4
be	1	#####	2	20	19	1	19.5	0	2
bf	1	#####	2	15	19	-4	17.0	0	2
bg	2	#####	2	18	15	3	16.5	1	4
bh	2	#####	2	20	19	1	19.5	1	4
bi	2	#####	2	19	20	-1	19.5	1	4
bj	1	#####	2	13	15	-2	14.0	0	2
bk	2	#####	2	17	11	6	14.0	1	4
bl	1	#####	2	9	5	4	7.0	0	2
bm	1	#####	2	18	11	7	14.5	0	2
bn	1	0	2	13	9	4	11.0	0	2
bo	1	3	2	19	18	1	18.5	0	2
bp	1	0	2	16	13	3	14.5	0	2
bq	1	3	2	11	11	0	11.0	0	2
br	1	0	2	17	18	-1	17.5	0	2
bs	1	0	2	13	11	2	12.0	0	2
bt	1	0	2	17	15	2	16.0	0	2