Lancaster University Management School

Economics Working Paper Series

2022/008

The effect of test anxiety on high stakes exams: evidence from an RCT using biomarkers

Emanuela Macrí and Giuseppe Migali

The Department of Economics
Lancaster University Management School
Lancaster LA1 4YX
UK

The effect of test anxiety on high stakes exams: evidence from an RCT using biomarkers

Emanuela Macri'*

Giuseppe Migali†

November 11, 2022

Abstract

We run a randomized control trial in an Italian university to study the effect of test anxiety on a high stakes exam. We separate students in two groups and we expose them to two random treatments, silence and music, that influence their level of pre-test anxiety. We measure the variation of test anxiety by observing the difference in individual biomarkers collected before and after the treatments. We find that a reduction in the mean arterial pressure and systolic pressure improve females test scores, and the effect is much stronger if the treatment is silence. For males we do not find any significant effect. Hence, we conclude that test anxiety may help to explain gender differentials in performance.

JEL Classification: I19, I20, I21.
Keywords: Test anxiety, biomarkers, RCT, high stakes exam, gender difference.

*Dipartimento G.E.S., University Magna Graecia, Catanzaro, Italy.

†g.migali@lancaster.ac.uk, Department of Economics, Lancaster University Management School, UK, and Dipartimento G.E.S., University Magna Graecia, Catanzaro, Italy. We thank Colin Green, Vincenzo Carrieri and the participants to SIEP 2020 conference for useful comments.

1 Introduction

Anxiety is a major factor that affects academic performance of students worldwide. Indeed, recent surveys of the American College Health Association (ACHA-NCHA, 2015, 2019) show that on average 27% of students report that anxiety may have a negative effect on their grades. A PISA study (OECD, 2017) shows that on average across OECD countries 55% of students are very anxious before a test even if they are well prepared. Furthermore, in recent years we have observed, in many countries throughout the world, an increase in the use of standardized tests for high-stakes decisions regarding students. In the UK, students typically starting from the age of 16 take several substantial examinations. The US often requires students to take at least one standardized test per year. In China there is a "higher examination", a college entrance exam which is extremely high pressure. Other examples of high stakes exams are in Japan, Canada, Mexico and most European countries. Furthermore, universities are increasingly relying on standardized tests of English proficiency to enroll international students, and this selection process is considered high-stakes (Cross and O'Loughlin, 2013). This naturally leads to concerns about increases in test anxiety and effects on students achievements.

In our work, we aim to provide the first causal evidence on the effect of test anxiety on high stakes exam performance. To do this we generate an exogenous variation in pre-test anxiety, exposing students to two treatments, that previous research suggests have differential effects on reducing anxiety, silence and music.[1] Our second contribution is the investigation of the

[1] The effect of music on academic performance has been widely investigated in the literature. The seminal work of Rauscher et al. (1993) finds an improvement in students test performance, the so called "Mozart Effect", whereas other studies fail to replicate the same results as documented by Steele et al. (1999a,b). However, there is only a small literature that suggests that music reduces test anxiety e.g. Haynes et al. (2004); Sezer (2009); Goldenberg et al. (2013).

potential links between test anxiety and individual biomarkers. The positive association between anxiety and blood pressure[2] is well known in the medical literature (Conley and Lehman, 2012; Wright et al., 2014; Ifeagwazi et al., 2018).

We run a randomised control trial on a sample of first year undergraduate law students, required to take a compulsory end of course exam in Public Economics. In our experiment, we measure students blood pressure (BP) and heart rate (HR) on two occasions, before and after exposing students for 15 minutes to two types of treatment, silence and music. We find that silence has a bigger effect than music in reducing students' test anxiety, as measured by a decrease in their systolic blood pressure and mean arterial pressure (MAP), after treatment. These effects are particularly relevant for females, indeed lower test anxiety improves their test scores by 11%, and when restricting to the silent group females scores increase by 17%. We do not find any statistically significant effect for males. Hence, our findings confirm the important role that test anxiety may have in generating gender differentials in high stakes exams.

The structure of the paper is the following: in the next Section we discuss the recent literature, in Section 2 we illustrate the experimental design and the econometric approach. Section 3 shows the results and the robustness checks. Section 4 concludes.

1.1 Our work and recent literature

The long run economic and social effects of underperforming in exam are a prominent focus for research. Indeed, the achievement of particular exam thresholds, that lead to key qualifications, is considered essential by educators, employers and governments. As shown by

[2] Systolic blood pressure is one of the physiological measures of anxiety.

Machin et al. (2020) failing a high stakes exam at lower secondary school may have important consequences in the completion of the studies (e.g. to drop out) and increases the probability to become a NEET (not in education, employment or training). Moreover, there is a growing body of studies that examine gender differences in performance and shows that females on average underperform in a stressful environment, where they usually do not succeed, and seek to avoid competitive situations (Gneezy et al., 2003; Niederle and Vesterlund, 2007).

Ors et al. (2013) as well as Jurajda and Münich (2011) find bigger gender differences in favour of men when the competitiveness of the examination increases. Pekkarinen (2015) analyses the results of the entrance examination in economics and business administration at Finnish universities, finding that, after controlling for starting points, women perform worse than men and are less likely to gain entry. De Paola and Gioia (2016) run a randomized control trial (RCT) on a sample of undergraduate students in economics and find that females under time pressure have worse test scores. A study from Ballen et al. (2017) conducted at Minnesota university on a sample of biology students shows that females perform worse than men in high stakes exams. Schlosser et al. (2019), using an experimental setting, contrast the performance of different demographic groups on low stakes and high stakes situations. They find that females and minorities outperform males and whites, respectively.

Nevertheless, it is important to understand the mechanisms behind these gender differences in performance. One possibile explanation is provided by the tournament theory: risk taking behaviours and overconfidence may determine gender differences in the choice of a strategy, subsequently reflected in the performances (Hvide, 2002; Goel and Thakor, 2008). Another important driver of gender difference is test anxiety, which has detrimental effects on the academic performance of many university students, especially females. There is a large literature outside economics, mainly developed by educationalists and psychologists, that examines test anxiety and shows its negative correlation with test performance (see meta analysis of Hembree, 1988; Seipp, 1991; Putwain and Best, 2011; Bonaccio et al., 2012; Putwain and Symes, 2018). Other studies demonstrate that females suffer test anxiety more than boys, and this is negatively reflected in their academic achievements (Bandalos et al., 1995; Ackerman and Heggestad, 1997; Peleg-Popko et al., 2003; Else-Quest et al., 2012; Putwain and Daly, 2014; Núñez Peña et al., 2016; Brandmo et al., 2019). Most of these works are descriptive and mainly based on correlations.

Furthermore, in our work, the use of biomarkers is instrumental in demonstrating the effect of silence/music in reducing test anxiety, and consequently its causal effect on test scores. The definition of test anxiety is commonly based on three items (experience of worry, tension and bodily symptoms) obtained from self-reported surveys, where a group of participating students is required to complete a questionnaire.[3]

There are few studies that directly link test anxiety and biomarkers. In the educational context, Zeller et al. (2004) observe 121 medical students taking the final licensing exam and find a positive change in their diastolic BP before and after the exam. Marazziti et al. (2007), considering 22 medical residents, show that systolic BP and HR are significantly higher when subjects are sitting before an examination, as compared to a calm situation. Moreover, they find a positive correlation between BP and self-assessed test anxiety defined using the Hamilton rating scale for anxiety (Hamilton, 1959) .

Zhang et al. (2011) evaluate, for few days, the BP and heart rate change of 64 medical

[3]Questionnaires are often structured using either the Spielberger Test Anxiety Inventory (Spielberger, 2010) or the reactions to test questionnaire (Sarason, 1984) or the Cognitive Test Anxiety Scale (Cassady and Johnson, 2002).

college students and on the first day of the exam period they ask students to complete the Zung Self-Rating Anxiety Scale (SAS) questionnaire. They find a significant increase in the systolic and diastolic BP, from the pre-exam to the exam period, and this change is positively correlated with the SAS score.

Conley and Lehman (2012) consider 99 undergraduate psychology students and monitored their BP on five consecutive days, and at the end of each day students were required to complete a survey used to define test anxiety. They discover a positive link between test anxiety identified in the survey and systolic BP as a response to academic stressors, such as examinations.

Finally, to the best of our knowledge, this is the first paper in economics that studies simultaneously the relationship between silence/music, test anxiety and test scores in an experimental setting. Indeed, Powdthavee (2010) estimates the effect of biomarkers on compulsory schooling in a non-experimental setting, using data obtained from the Health Survey for England and not focusing on test anxiety. Carrieri and Jones (2017) employ the same data but they are mainly interested in explaining gender differentials in biomarkers in England. Goldenberg et al. (2013) run an RCT on a sample of 176 college undergraduate psychology students who, before taking an exam, are randomly assigned to two groups, one listening Mozart's music and the other remaining silent. They find no significant group differences in the mean levels of test anxiety and those exposed to music have lower test scores. However, although the structure of this study is very similar to ours, they only measure test anxiety by exploiting a post-exam questionnaire and not using biomarkers. Lilley et al. (2014) find that calm music, compared to obnoxious music, reduces systolic BP and HR levels and has positive effects on test scores. They also assess test anxiety using a self-reported measure based on the participant's score on the Spielberg test anxiety inventory and they find a positive correlation with the biomarkers. Nevertheless, they do not perform any randomisation.

2 Methodological approach

2.1 Experimental Design

In our analysis we have selected all first year undergraduate students enrolled in the Law Degree at Magna Graecia University (MGU) and required to take the high stakes exam in Public Economics, a compulsory course usually taught during the second semester. Furthermore, this is the only economic exam in the Law degree, if students do not pass it cannot graduate. Taking into account that most students do not have an economic background, this exam puts a lot of pressure on them and can be considered more high stakes than others.[4]

Two weeks before the end of the lectures, students are informed about the possibility to participate in an empirical investigation that involves the measurement of their blood pressure and heart beat. Participation is on voluntary basis, and students were aware in advance of the biomarkers measurement and signed a consent form. So there is no surprise effect. To avoid possible self-selection among participants we do not reveal to students, and to anybody else at the university, that we are planning a field experiment.[5]

[4]In the Italian system all exams are high stakes, since students must pass all exams and no single fail is allowed/condoned in order to get a degree. First year exams count as much as any following year ones. In fact the final graduation mark will be given by a combination of the rescaled average mark in all exams taken during the course of study and the dissertation mark. Students have very high incentives in taking the exam of Public Economics in the form proposed by our experiment because it's relatively easier than a standard exam. As shown in Table 1, when the exam is a multiple choice test, the pass percentage is over 75%. Furthermore, although students are allowed to retake an exam many times, if they don't graduate within the statutory end of the degree course, they must re-enroll each year paying the full annual fees.

[5]Our investigation has been undertaken in an academic year prior to the COVID-19 pandemic, so our results are not affected in any way.

We have offered strong incentives in terms of exam structure and programme, and the university ethic committee has approved the experiment. In particular, students have the opportunity to answer a multiple choice test limited to the topics taught during the lectures. This reduction corresponds to two thirds of the full programme. Therefore, we are confident we can exclude the risk taking behaviour as driver of our results, leaving test anxiety as the main mechanism that could explain the gender difference in performance.

The test lasts 1 hour and includes 15 questions for a total of 31 points.[6] The multiple choice questions exam format has been chosen for its objectiveness in grading, and the test has been scientifically validated by an expert affiliated to a well known Italian university. The experiment is set at the first exam session after the end of the course, i.e. the summer session in June.[7] The potential number of participants is 233, all first year Law students, and 173, around 74%, decided to join the experiment. Due to university regulations we have to admit additional 13 students enrolled in the second or following years, hence, we have a total of 186 participants. However, 2 students quitted the exam before the start and are excluded from the analytical sample (184 observations). For students not participating in the experiment, and sitting any future session, the standard exam lasts 1 hour and consists in three open questions that could cover any topic detailed in the syllabus. Then, the first available standard exam date is the day after the experiment, only 11 students attended that session, and 2 passed the exam. In the following 2 exam sessions, there were 39 students (7 passed the exam) and 11 students (3 passed the exam), respectively.

Since the course material and the lecturer in Public Economics have been unchanged for

[6] Each question carries two points, except the last which carries 3 points. The grades in the Italian academic system range from 0 to 30 summa cum laude, which corresponds to 31.

[7] In Italy, most universities offer several exam sessions, at MGU there are usually 7 per academic year.

the past few years, in Table 1 we make some comparison with the students behaviour and performance in the past. In the first exam session in June, we usually register the highest number of enrolled[8] and in the year of the experiment there is a clear increase (see column 2). Furthermore, it is evident that the structure of the exam affects the percentage of pass. A multiple choice test is relatively more affordable for the students,[9] in fact with this structure the percentage of pass is above 75% (see column 4, Table 1), whereas with an open questions exam the percentage drops by around 20 percentage points. Hence, we are confident we are giving students the correct incentives to maximise attendance at the experiment and avoiding risk taking behaviours or tempting the fates.

Treatments

Students are exposed to two types of treatment that, on average, are expected to reduce test anxiety: music and silence. Indeed, classical music can create a calming effect on the listener. This is due to the release of dopamine which is the body's natural happy chemical that improves a person's mood, and also blocks the release of stress (Ferreri et al., 2019). There is a variety of activities that release dopamine and listening to classical music is one of them.

Ideally, we would need a suitable control group to generate a counterfactual situation for the music, and one possibility could be the "business as usual" situation, i.e. leaving students without any treatment. Unfortunately, in the Italian university system students are usually allowed to enter the examination room before the exam. They can sit anywhere, and

[8]The majority are first year students, although there could be students of the following years.

[9]Since some students were tempting fate attending several sessions (without limitations) until they passed the exam, the structure has been changed over the years from multiple choice to open questions.

may talk, rehearse, argue, shout, use a mobile phone, a laptop or any other electronic tool. Once the examiner is ready, students may be re-allocated to different seats, are required to leave all their belongings in a corner of the room and are provided with the exam scripts. They must remain silent only for the duration of the test. Consequently, a "business as usual" situation may be a source of anxiety as well as relaxation, or other emotional statuses (e.g. getting nervous, annoyed, distracted). All these reactions are impossible to control and identify in an experimental setting and they may in turn affect the exam performance. Consequently, to have a valid alternative to the music we have decided to introduce silence as a second treatment. We ask students to remain silent for a time period exactly equal to the music treatment.

The choice of silence is also supported by the literature, indeed, a study of Bernardi et al. (2006) show a more evident relaxing effect of silence, in terms of reduced heart rate and blood pressure, compared to slow or meditative music. Recently, Pfeifer et al. (2016) demonstrate that music-induced relaxation is more effective when combined with silence; Malakoutikhah et al. (2020) find that different genres of music, compared to silence, do not have extra effect in reducing anxiety and improving relaxation. In the educational context, Lai et al. (2008) show that both silence and music lessen students' examination anxiety.

Experiment day

Participants to the experiment are registered and randomized. Afterwards, they are separated in two rooms of the same dimension, located on the same floor and with a constant temperature (20 degree Celsius). In a first room students are randomly allocated to specific seats and required to listen a Mozart symphony for 15 minutes. In a second room, students are also randomly allocated to specific seats and required to remain silent for 15 minutes. Experts from the Faculty of Medicine from MGU[10] measures twice, before and after each treatment, the blood pressure (BP) and the heart rate (HR)[11]. The values collected for each measurement are not disclosed to the students until the end of the experiment. Therefore, they could have not directly affected their test anxiety. These biomarkers are considered objective indicators of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention.[12] As discussed in Section 1.1, they are also employed to measure test anxiety. After the collection of the second medical measurement, students can start their exam. They are free to withdraw at any time, if they have second thoughts about taking the exam under the conditions that are part of the experiment. At the end of the exam, students complete a questionnaire regarding their family background, eating habits, health status and cultural habits. The exam script, the medical measurements, the questionnaire and the authorisation to personal data processing are included in a sealed envelope which has an identification number. The anonymised exams scripts are graded by a pool of 3 lecturers, who give a score ranging between 0 and 31.

2.2 Data Preparation and descriptive statistics

We use detailed administrative data, provided by MGU, and we stratify students by: gender, high school grade (divided in quartiles), domicile in the same town where the university is

[10]One for every 10 law students and accompanied by a MD supervisor. Furthermore, the Department of Medicine has borrowed (from a national provider) several new and identical blood pressure gauges, for the purposes of our experiment.

[11]Students are required to remain sit before and after any measurement and during the treatment.

[12]This definition is given by the National Institute of Health Biomarkers Definitions Working Group, see Atkinson et al. (2001).

located, eligibility to scholarship. We obtain 26 strata, and students included in each strata are randomly allocated to one of two treatment groups.

Table 2 reports some descriptive statistics, for the full sample of 184 students included in the experiment. The average age is around 21 years old, 70% of students are females[13], and about 20% resident in the university town. Students have an average high school grade of 83 out of 100, 70% of them attended a lyceum and only 24% is eligible for a scholarship. The percentage of students with a graduate father or mother is around 21-22%. The majority of students (93%) has a working father, whereas less than 50% has a working mother. Finally, about 31% are smokers, and in terms of leisure activities around 50% used to go to concerts, listen to some music and use social network.

In Table 3 we report the balance checking of our randomisation. We compare the averages of the observed individual characteristics in the two groups, and report the t-tests of equality of means. Overall, we do not notice any significant difference between students in the silence and music groups, especially in the variables used for the stratification (gender, residence, lyceum, scholarship eligibility and high school grade). Thus, we are confident of a successful randomisation. At the bottom of Table 3 we show the average test score, post-treatment, in the two comparison groups and as expected the difference is statistically significant, although only at 10%.

One of the main advantages of our analysis is the use of biomarkers to measure the variation of test anxiety before and after the treatment. We collect systolic and diastolic pressure (see Figure 1) and we use this information to compute the mean arterial pressure

[13]This is slightly above the annual average (around 62%) at both at MGU and at national level, see Panel A Table A1 in the appendix.

(MAP). This is a measure of the average BP in an individual during a single cardiac cycle, it normally ranges between 65 and 110 mmHG and is easily approximated by the following formulae: *MAP = diastolic BP + 1/3 (systolic BP - diastolic BP).*

Figure 2 displays the distribution of the MAP variation by treatment status and gender. For females, it is evident that silence generates a big negative median MAP variation (measured in mm Hg), whereas the median variation in the music group is around zero. For males, the effect of both treatments is less strong, in fact, in the silent group the median MAP variation is around zero and it is slightly negative in the music group.

In Table 4, we show all biomarkers measurements collected before and after any of the two treatments, together with the MAP. In panel A, we notice that, for all students, all biomarkers are reduced after treatment. Comparing students by gender, it is clear that the most pronounced reductions are registered for females, especially with systolic BP and MAP (see Panel B). We also report the heart beat, for illustrative purposes, and we notice a reduction after treatment.[14] In Table 5, focusing on the latter two biomarkers, it is obvious that silence has a bigger effect than music in reducing test anxiety. Precisely, in panel A, for students in the silent group we observe a 2% reduction of systolic BP and MAP, the double compared to students in the music group. For females, in particular, remaining silent generate a 3% reduction in both biomarkers (see Panel B). Males are more affected by music, although only systolic BP has relevant variation (-3%).

Finally, in Figure 3, observing how the two treatments indirectly affect students' test performances, it is confirmed the stronger effect of silence in reducing females' test anxiety.

[14]In the econometric analysis we only focus on the blood pressure because we do not find statistically significant effects using the heart beat as an alternative measure of test anxiety.

We notice that for males the median grade is similar in both comparison groups (21 out of 31), although the grade distribution is less concentrated in the music group. If we compare females and males in the music group, we observe that females have a lower median grade (19 vs 21, respectively). Whereas females in the silence group perform better than males (23 vs 21, respectively). Females in the silence group have a higher median grade compared to females in the music group (23 vs 19).

2.3 Sample restrictions

From our initial sample size of 184 students, we make further restrictions using administrative information on students' characteristics and the post-experiment questionnaire. We exclude students who have reported illnesses,[15] that could affect the measurement of the heart rate and the blood pressure, precisely, hypertension, hypotension, hearth diseases, diabetes, hyperthyroidism and kidney failure. Following the American Heart Association guidelines (see Figure 1) BP numbers of less than 120 mmHg (systolic) and 80 mmHg (diastolic) are considered within the normal range. For this reason we also exclude students who have abnormal biomarkers, precisely HR above 110, diastolic BP above 100 mmHg and systolic BP above 140 mmHG. We also drop one student that left the experiment before the conclusion and obtained a score of 8. Our final sample size includes 148 observations.

[15]Since this information has been collected after the exam, it was not available for the randomisation. This is another reason for the exclusion.

2.4 Econometric approach

The estimation of a linear model of test anxiety on test scores would be affected by several sources of endogeneity. Indeed, test anxiety can increase for students who expect to do worse on the exam (reverse causation), for those who had a shock to their preparation beforehand, and somehow students more/less anxious do better/worse on average on tests. Hence, if we simply define the students' test anxiety by measuring their biomarkers before the exam, the effect on test scores would be biased. We therefore expose students to two random treatments, silence and music, to generate an exogenous variation in their test anxiety. We measure this variation by computing the difference in their mean arterial pressure (MAP), before and after the treatments.

As noticed in Figure 2 and Table 3, on average the MAP decreases after both treatments, however for some students it may actually increase and the variations are clearly different for males and females. Since our interest is on the evaluation of the effects of a reduction of test anxiety, to better interpret our results, we define a dummy variable, called MAP decrease (Δ MAP^{-}) equal to 1 if a student has a reduction in his MAP after any treatment and 0 if she has an increase. In Panel D of Table 5 we report the total number of students, by gender, classified as having a MAP decrease, and we observe a similar proportion of males and females (over 61%).

Thus, we estimate the following model, separately for males and females

$$Y_i = \alpha + \beta_1 \Delta MAP_i^- + \beta_2 X_i + \epsilon_i \quad (1)$$

where Y is the log of the exam scores, and the vector X includes additional controls on individual and family characteristics, such as age, eligibility for a scholarship, father with HE degree and cramming (rehearsal). We also add some variables that inform on individual habits such as listening music, going to concerts and using social networks. β_1 is our coefficient of interest: if students exposed to any of the two treatments have a reduction in their test anxiety, we expect a positive and statistically significant effect on their test scores.

To disentangle the specific effect of each treatment on the reduction of test anxiety, we first estimate their direct effect on the MAP variation. Then, to establish the indirect effect on test scores, we re-estimate equation 1 restricting to the subsamples of students in the silent group and the music group. This would be enough to determine the impact of test anxiety on test scores, since students are randomised into groups.

Furthermore, to assess the validity of our results we run a series of robustness checks. We first employ for the analysis a different biomarker, replacing the reduction of MAP with a reduction of systolic pressure, a well-known measure of anxiety. Second, we estimate a reduced form model of the effect of each treatment on test scores

$$Y_i = \alpha + \gamma_1 S_i + \gamma_2 X_i + \epsilon_i \tag{2}$$

where S_i is a dummy variable equal to 1 if a student is included in the silent group and 0 if he is in the music group. X includes the same covariates as in equation 1. If students exposed to silence have a higher reduction in test anxiety, we should expect the coefficient γ_1 to be positive and statistically significant. Finally, for illustrative purposes, we attempt to estimate the average treatment effect of test anxiety on test scores, by highlighting the role of the treatments as possible mediators. To do so we adopt an inverse probability weighting risk adjusted model. Due to the small sample size, but confident in the fact that students have already been randomised, we do not include any covariate in this model.

3 Results

Table 6 shows the main results of our analysis, obtained estimating equation 1, where all students are exposed to one of the two treatments.[16] Contrasting the effects of a MAP reduction for females (col. 1-3) with males (col. 4-6), it is immediately clear that a reduction of test anxiety has a positive and statistically significant effect on test scores only for females.[17] The size of the effect is around 13% in a model without any covariate. When we add individual controls, such as age, eligibility for a scholarship and father with a degree, the magnitude of a MAP decrease is slightly smaller, 11%. Adding cramming and leisure activities the effect is around 10%, which corresponds to around 3 grade points out of 30. In this first set of results, we are not fully exploiting the advantages of the randomisation, indeed the effect of a MAP reduction is sensitive to the inclusion of covariates.

To evaluate the specific effect of each treatment on the reduction of test anxiety, we can observe in Table 7 that for students in the silent group the probability to have a reduction of the MAP increases by around 19pp. For males, it appears that being in the music group has a more relaxing effect, however the effect is not statistically significant. The results are unchanged when adding control variables.

[16]We have also estimated equation 1 in the full sample including an interaction between MAP variation and sex. We have found a stronger effect for females, however the estimates are less precise. Hence, we prefer to report the results separately for males and females.

[17]The smaller sample size of the male group might affect the precision of our results. However, as reported in Panel A of Table A1 males are around 38% of the students enrolled in Law, both at national level and at MGU.

In Table 8, we report the results obtained estimating equation 1 restricting to the subsamples of females included in the silent group (col. 1-3) and in the music group (col. 4-6). As already established before (see Figure 2, Table 5 and Table 7), for females silence has a stronger effect than music in reducing test anxiety, and this is also reflected on test scores. In fact, a reduction of the MAP increases test scores by 17% in the silent group, consistently across all specifications. Whereas, for females in the music group the effect is still positive but smaller in magnitude and not statistically significant. For males the effects are not statistically significant, confirming that test anxiety is mainly affecting females performances.[18]

3.1 Robustness checks

To assess the validity of our findings we undertake a series of robustness checks.

One possible drawback in our analysis could be that the initial level of anxiety could affect the reactivity to each type of treatment. Knowing that the allocation to each treatment is random between two balanced groups, and students are not aware of the type of treatment during the first measurement of their BP, we have run a regression of each pre-treatment BP measurement directly on test scores. We do not find any statistically significant effect, both for males and females. We have also disaggregated the pre-treatment MAP in quartiles and run a regression on test scores and we still do not find any significant direct effect, at any quartile. Moreover, we have compared the MAP variation by quartile of pre-treatment MAP and by treatment status. We do not notice any differential response to treatment by quartile of ex-ante anxiety.[19]

[18]The results are available upon request.
[19]The results are available upon request.

Our second robustness check is to re-estimate equation 1 using a different biomarker. Following the medical literature (e.g. Conley and Lehman, 2012; Wright et al., 2014), we know that the systolic BP is a common measure of anxiety, and we use it to replace the MAP to evaluate the variation in test anxiety. We have first replicated the estimates reported in Table 7, replacing the MAP variation with the systolic BP variation, as dependent variable. The results[20] confirm that for females included in the silent group, the probability to have a reduction in their systolic BP increases by around 20pp. For males the effect is still not statistically significant.

In Table 9, col.1-3, we restrict the sample on females only, and we observe that a reduction of the systolic BP induces a positive and highly statistically significant increase in test scores, by around 16.5%, consistently in all specifications. The same results are confirmed when we restrict the analysis only to females included in the silent group (col.4-6): the increase in test scores due to a reduction of systolic BP is, indeed, around 17%. For males, the use of a different biomarker does not produce any significant effect on test scores.

Third, we estimate reduced form models of the effect on test scores of being included in the silent group or in the music group. In Table 10, we contrast the findings for females (col.1-3) with those for males (col.4-6), and we observe a positive and statistically significant effect only for females. Precisely, being exposed to silence increases females test scores by around 12-13% (depending on the specification) more than being exposed to music.

Hence, our results confirm that silence has an important role in reducing test anxiety and indirectly improving females grades. To shed further light on the described mechanism, we estimate, for illustrative purposes, a simple inverse probability weighting risk adjusted

[20] Available upon request.

model,[21] focusing only on females. We want to evaluate the causal effect of a reduction of MAP on test scores, using silence as possible mediator. Looking at Table 11, Panel 1, we find that the average treatment effect of a MAP reduction corresponds to a 12% increase in test scores. In Panel 2, we show the outcome models, where the effect on test scores of being exposed to silence, is estimated for females that register either an increase or a decrease in their MAP. We notice that silence has a positive and significant effect on test scores, by around 14%, only if there is a reduction in test anxiety (i.e. MAP reduction). Panel 3 confirms, as already seen in Table 7, that silence is reducing test anxiety, in fact for females included in the silent group there is a 78% probability of having a MAP reduction.

Overall, our robustness checks confirm that test anxiety is a relevant mechanism in explaining the gender differential in a high stakes exam, and if we manage to exogenously reduce test anxiety females have a strong improvement in their grades.

4 Conclusion and Discussion

In this paper we evaluate the effect of test anxiety on academic performance. We generate an exogenous variation in pre-test anxiety by exposing a sample of first year students in law to two treatments, that are known to reduce anxiety. Before taking a compulsory end of course exam in Public Economics, the students are randomised in two groups, one required to remain silent and the second group exposed to classical music. We measure the variation

[21] As mentioned in Section 2.4, we are not able to include covariates in this model. Furthermore, we have also performed an IV strategy instrumenting the MAP variation with treatments, and overall our results are confirmed for females, and we do not find any effect for males. In the first stage, silence has a positive and statistically significant effect (at 10%) on the probability of having a MAP decrease. The F-test passes but is relatively low. In the second stage a MAP reduction increases test scores, however the effect is marginally significant and relatively high in magnitude, possibly because the IV is estimating a LATE. The results are available upon request.

in test anxiety by collecting individual biomarkers before and after each treatment. We find that silence has a stronger effect than music in reducing anxiety, as shown by the reduction of the mean arterial pressure and systolic blood pressure. The effect of less test anxiety is particularly relevant for females, indeed we find an increase in their test scores by around 12% and when restricting to the females in the silent group the effect is even stronger, around 17%.

We have already seen from the literature in the field that females are the most affected by test anxiety and tend to under-perform. However, that does not imply that females' performances are worse than males' performances, our point is that females could have performed much better without anxiety. To have an idea of past performances in the same exam by gender, we have got additional administrative data from MGU, that we reported in Panel C of Table A1. We compare females and males in the first exam session (June) in the 3 years before the experiment, and we observe that females on average perform slightly better than males. However, since we do not find any significant effect on males in our experimental setting, we argue that test anxiety can be considered as an important driver of the gender differentials in performance, especially in high stakes exams.

It is worth mentioning some possible caveats concerning the external validity of our findings. As participants in our experiment are from a medium university in the South of Italy, they usually come from more disadvantaged socio-economic background and tend to perform slightly worse than students from the North (e.g. OECD, 2009). However, that means that students feel more the pressure of getting a degree and all exams can be considered high stakes. Using the information available from USTAT-MUR (2022) on the population of first year enrolled in Law in Italy, we observe that, in the academic year of the experiment and the three years before, the students' composition at MGU mirrors well the national proportions. Around 60% of the first year students in Law are females (see Panel A and B Table A1), and we notice a decreasing trend in the enrolments. Hence, the larger and statistically significant effects that we find only for females might be extended at least to other universities of the same size of MGU.

Overall, possible explanations of the observed gap can be related to females' higher emotionality (Mueller and Courtois, 1980; Else-Quest et al., 2012), they could consider the situations in which they must be judged as fertile ground for the development of gender stereotypes (Osborne, 2006). In particular, subjects like mathematics and science could generate what psychologists call a stereotype treat, "math is for boys", which get girls more anxious than boys. Females may have threatening rather than challenging perception of evaluative situations, due to self-doubt regarding the ability to cope with an exam (Jerusalem and Schwarzer, 1992; Cassady and Johnson, 2002), and these fears are transformed into test anxiety. Consequently, policymakers need to bear in mind the gender differences in the effects of test anxiety, knowing that may ultimately affect the females outcomes later in life.

Furthermore, specific educational policies could be implemented to reduce test anxiety, taking into account that direct knowledge of their anxiety status might hurt students, as shown by Azmat et al. (2019) in their study on feedback on performance. For example, it could be more helpful to provide counseling services and psychotherapy, as well as implementing anxiety management programs. Differentiated kinds of counseling and psychological training, for female and male students, could help to reduce anxiety before examinations. The existing programs should be strengthened because, as shown by Cage et al. (2020), there are several barriers that prevent students from accessing support, such as current mental health conditions (depression, anxiety and stress). Possible solutions that could increase the welfare for the students, as suggested by Walsh et al. (2022), are the reduction of waiting times and the provision of more information about the counseling services, specifically to first year students or those less-privileged. In particular, preventive programs for first-year students, tackling test anxiety, could be introduced in the context of the transfer from school to university, which is particularly challenging (Laidlaw et al., 2016). Overall, policymakers need to bear in mind the gender differences in the effects of test anxiety, knowing that may ultimately affect the females outcomes later in life.

References

ACHA-NCHA, A. C. H. A. (2015). American college health association-national college health assessment ii: Reference group undergraduates executive summary fall 2015,2016,2017,2018,2019. *Hanover, MD: American College Health Association.*

ACHA-NCHA, A. C. H. A. (2019). American college health association-national college health assessment ii: Reference group undergraduates executive summary spring,2016,2017,2018,2019. *Hanover, MD: American College Health Association.*

Ackerman, P. L. and Heggestad, E. D. (1997). Intelligence, personality, and interests: evidence for overlapping traits. *Psychological bulletin*, 121(2):219.

Atkinson, A. J., Colburn, W. A., De Gruttola, V. G., DeMets, D. L., Downing, G. J., Hoth, D. F., Oates, J. A., Peck, C. C., Schooley, R. T., et al. (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical pharmacology & therapeutics*, 69(3):89–95.

Azmat, G., Bagues, M., Cabrales, A., and Iriberri, N. (2019). What you donÂt knowÂ canÂt hurt you? a natural field experiment on relative performance feedback in higher education. *Management Science*, 65(8):3714–3736.

Ballen, C. J., Salehi, S., and Cotner, S. (2017). Exams disadvantage women in introductory biology. *PLoS One*, 12(10):e0186419.

Bandalos, D. L., Yates, K., and Thorndike-Christ, T. (1995). Effects of math self-concept, perceived self-efficacy, and attributions for failure and success on test anxiety. *Journal of educational psychology*, 87(4):611.

Bernardi, L., Porta, C., and Sleight, P. (2006). Cardiovascular, cerebrovascular, and respiratory changes induced by different types of music in musicians and non-musicians: the importance of silence. *Heart*, 92(4):445–452.

Bonaccio, S., Reeve, C. L., and Winford, E. C. (2012). Text anxiety on cognitive ability test can result in differential predictive validity of academic performance. *Personality and Individual Differences*, 52(4):497–502.

Brandmo, C., Bråten, I., and Schewe, O. (2019). Social and personal predictors of test anxiety among norwegian secondary and postsecondary students. *Social Psychology of Education*, 22(1):43–61.

Cage, E., Stock, M., Sharpington, A., Pitman, E., and Batchelor, R. (2020). Barriers to accessing support for mental health issues at university. *Studies in Higher Education*, 45(8):1637–1649.

Carrieri, V. and Jones, A. M. (2017). The income-health relationship beyond the mean: New evidence from biomarkers. *Health economics*, 26(7):937–956.

Cassady, J. C. and Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary educational psychology*, 27(2):270–295.

Conley, K. M. and Lehman, B. J. (2012). Test anxiety and cardiovascular responses to daily academic stressors. *Stress and Health*, 28(1):41–50.

Cross, R. and O'Loughlin, K. (2013). Continuous assessment frameworks within university english pathway programs: Realizing formative assessment within high-stakes contexts. *Studies in Higher Education*, 38(4):584–594.

De Paola, M. and Gioia, F. (2016). Who performs better under time pressure? results from a field experiment. *Journal of Economic Psychology*, 53:37–53.

Else-Quest, N. M., Higgins, A., Allison, C., and Morton, L. C. (2012). Gender differences in self-conscious emotional experience: A meta-analysis. *Psychological bulletin*, 138(5):947.

Ferreri, L., Mas-Herrero, E., Zatorre, R. J., Ripollés, P., Gomez-Andres, A., Alicart, H., Olivé, G., Marco-Pallarés, J., Antonijoan, R. M., Valle, M., et al. (2019). Dopamine modulates the reward experiences elicited by music. *Proceedings of the National Academy of Sciences*, 116(9):3793–3798.

Gneezy, U., Niederle, M., and Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The quarterly journal of economics*, 118(3):1049–1074.

Goel, A. M. and Thakor, A. V. (2008). Overconfidence, ceo selection, and corporate governance. *The Journal of Finance*, 63(6):2737–2784.

Goldenberg, M. A., Floyd, A. H., and Moyer, A. (2013). No effect of a brief music intervention on test anxiety and exam scores in college undergraduates. *Journal of articles in Support of the Null Hypothesis*, 10(1).

Hamilton, M. (1959). The assessment of anxiety states by rating. *British journal of medical psychology*, 32(1):50–55.

Haynes, A. F., Mullins, A. G., and Stein, B. S. (2004). Differential models for math anxiety in male and female college students. *Sociological Spectrum*, 24(3):295–318.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of educational research*, 58(1):47–77.

Hvide, H. K. (2002). Tournament rewards and risk taking. *Journal of Labor Economics*, 20(4):877–898.

Ifeagwazi, C. M., Egberi, H. E., and Chukwuorji, J. C. (2018). Emotional reactivity and blood pressure elevations: anxiety as a mediator. *Psychology, health & medicine*, 23(5):585–592.

Jerusalem, M. and Schwarzer, R. (1992). Self-efficacy as a resource factor in stress appraisal processes. *Self-efficacy: Thought control of action*, 195213.

Jurajda, Š. and Münich, D. (2011). Gender gap in performance under competitive pressure: Admissions to czech universities. *American Economic Review*, 101(3):514–18.

Lai, H.-L., Chen, P.-W., Chen, C.-J., Chang, H.-K., Peng, T.-C., and Chang, F.-M. (2008). Randomized crossover trial studying the effect of music on examination anxiety. *Nurse Education Today*, 28(8):909–916.

Laidlaw, A., McLellan, J., and Ozakinci, G. (2016). Understanding undergraduate student perceptions of mental health, mental well-being and help-seeking behaviour. *Studies in Higher Education*, 41(12):2156–2168.

Lilley, J. L., Oberle, C. D., and Thompson Jr, J. G. (2014). Effects of music and grade consequences on test anxiety and performance. *Psychomusicology: Music, Mind, and Brain*, 24(2):184.

Machin, S., McNally, S., and Ruiz-Valenzuela, J. (2020). Entry through the narrow door: The costs of just failing high stakes exams. *Journal of Public Economics*, 190:104224.

Malakoutikhah, A., Dehghan, M., Ghonchehpoorc, A., Afshar, P. P., and Honarmand, A. (2020). The effect of different genres of music and silence on relaxation and anxiety: A randomized controlled trial. *Explore*, 16(6):376–381.

Marazziti, D., Ambrogi, F., Abelli, M., Di Nasso, E., Catena, M., Massimetti, G., Carlini, M., and Dell'Osso, L. (2007). Lymphocyte subsets, cardiovascular measures and anxiety state before and after a professional examination. *Stress*, 10(1):93–99.

Mueller, J. H. and Courtois, M. R. (1980). Test anxiety and breadth of encoding experiences in free recall. *Journal of Research in Personality*, 14(4):458–466.

MUR (2022). Open data. *Ministero Universita' e Ricerca.*

Niederle, M. and Vesterlund, L. (2007). Do women shy away from competition? do men compete too much? *The quarterly journal of economics*, 122(3):1067–1101.

Núñez Peña, M. I., Suárez Pellicioni, M., and Bono Cabré, R. (2016). Gender differences in test anxiety and their impact on higher education students' academic achievement. *Procedia-Social and Behavioral Sciences, 2016, vol. 228, p. 154-160.*

OECD (2009). Pisa 2009 assessment framework.

OECD (2017). *Pisa 2015 Results (Volume III): Students' Well-Being.*

Ors, E., Palomino, F., and Peyrache, E. (2013). Performance gender gap: does competition matter? *Journal of Labor Economics*, 31(3):443–499.

Osborne, J. W. (2006). Gender, stereotype threat, and anxiety: Psichophysiological and cognitive evidence. *Electronic Journal of Research in Educational Psychology*, 4(1):109–137.

Pekkarinen, T. (2015). Gender differences in behaviour under competitive pressure: Evidence on omission patterns in university entrance examinations. *Journal of Economic Behavior & Organization*, 115:94–110.

Peleg-Popko, O., Klingman, A., and Nahhas, I. A.-H. (2003). Cross-cultural and familial differences between arab and jewish adolescents in test anxiety. *International Journal of Intercultural Relations*, 27(5):525–541.

Pfeifer, E., Sarikaya, A., and Wittmann, M. (2016). Changes in states of consciousness during a period of silence after a session of depth relaxation music therapy (drmt). *Music and Medicine*, 8(4):180–186.

Powdthavee, N. (2010). Does education reduce the risk of hypertension? estimating the biomarker effect of compulsory schooling in england. *Journal of Human Capital*, 4(2):173–202.

Putwain, D. and Daly, A. L. (2014). Test anxiety prevalence and gender differences in a sample of english secondary school students. *Educational Studies*, 40(5):554–570.

Putwain, D. W. and Best, N. (2011). Fear appeals in the primary classroom: Effects on test anxiety and test grade. *Learning and Individual Differences*, 21(5):580–584.

Putwain, D. W. and Symes, W. (2018). Does increased effort compensate for performance debilitating test anxiety? *School Psychology Quarterly*, 33(3):482.

Rauscher, F. H., Shaw, G. L., and Ky, C. N. (1993). Music and spatial task performance. *Nature*, 365(6447):611.

Sarason, I. G. (1984). Stress, anxiety, and cognitive interference: reactions to tests. *Journal of personality and social psychology*, 46(4):929.

Schlosser, A., Neeman, Z., and Attali, Y. (2019). Differential performance in high versus low stakes tests: evidence from the gre test. *The Economic Journal*, 129(623):2916–2948.

Seipp, B. (1991). Anxiety and academic performance: A meta-analysis of findings. *Anxiety research*, 4(1):27–41.

Sezer, F. (2009). The effect of music therapy to reduction of high school studentsÃÂ examination anxiety. *e-Journal of New World Sciences Academy*, 4(3):859–871.

Spielberger, C. D. (2010). Test anxiety inventory. *The Corsini encyclopedia of psychology*, pages 1–1.

Steele, K. M., Bass, K. E., and Crook, M. D. (1999a). The mystery of the mozart effect: Failure to replicate. *Psychological Science*, 10(4):366–369.

Steele, K. M., Dalla Bella, S., Peretz, I., Dunlop, T., Dawe, L. A., Humphrey, G. K., Shannon, R. A., Kirby Jr, J. L., and Olmstead, C. (1999b). Prelude or requiem for the 'mozart effect'? *Nature*, 400(6747):827.

Walsh, S., Cullinan, J., Flannery, D., and Kennelly, B. (2022). Modelling student preferences for the design of campus counselling services. *Studies in Higher Education*, 47(2):305–317.

Wright, B. J., O'brien, S., Hazi, A., and Kent, S. (2014). Increased systolic blood pressure reactivity to acute stress is related with better self-reported health. *Scientific reports*, 4(1):1–5.

Zeller, A., Handschin, D., Gyr, N., Martina, B., and Battegay, E. (2004). Blood pressure and heart rate of students undergoing a medical licensing examination. *Blood pressure*, 13(1):20–24.

Zhang, Z., Su, H., Peng, Q., Yang, Q., and Cheng, X. (2011). Exam anxiety induces significant blood pressure and heart rate increase in college students. *Clinical and experimental hypertension*, 33(5):281–286.

Figure 1: Systolic and diastolic blood pressure levels

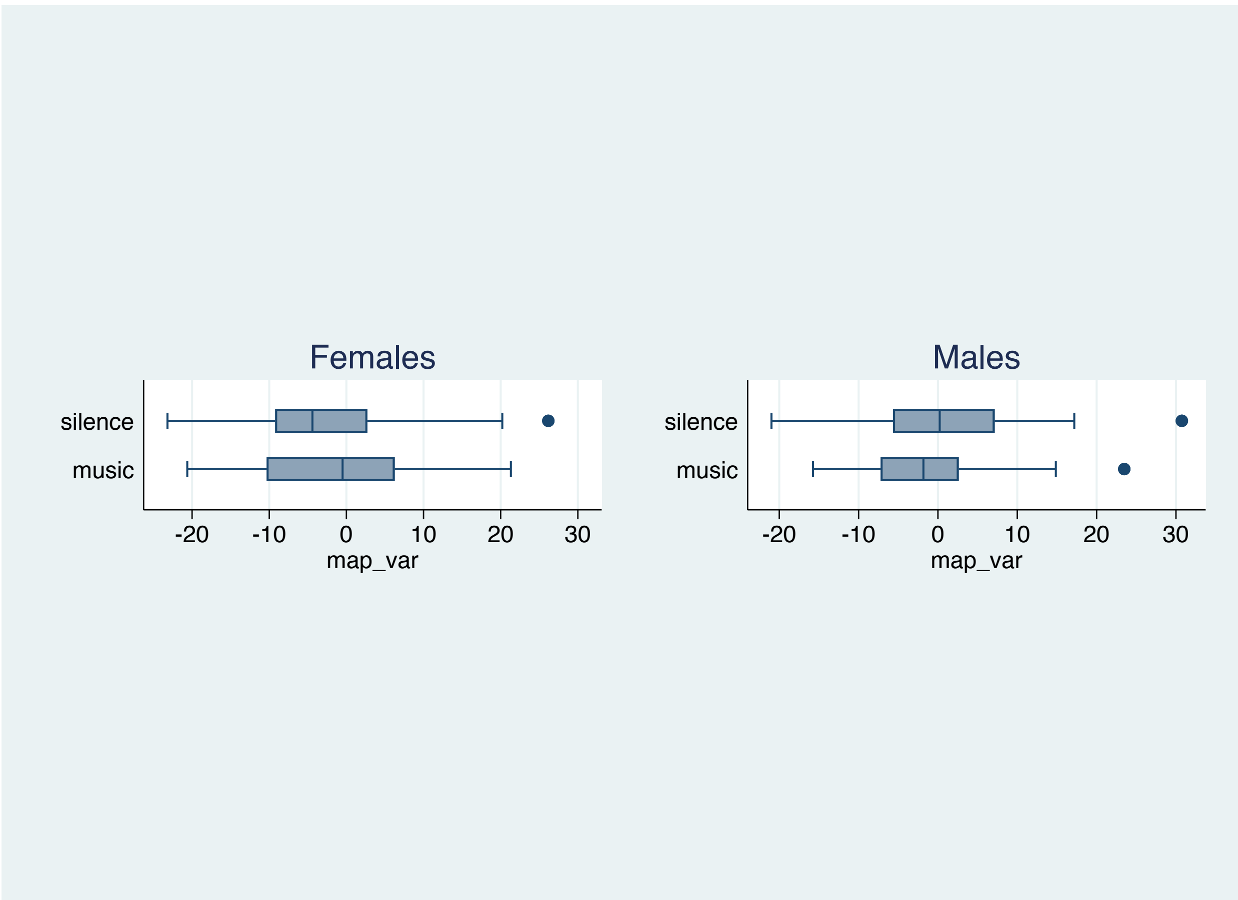
Blood Pressure Categories

American Heart Association | American Stroke Association

BLOOD PRESSURE CATEGORY	SYSTOLIC mm Hg (upper number)		DIASTOLIC mm Hg (lower number)
NORMAL	LESS THAN 120	and	LESS THAN 80
ELEVATED	120 – 129	and	LESS THAN 80
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 1	130 – 139	or	80 – 89
HIGH BLOOD PRESSURE (HYPERTENSION) STAGE 2	140 OR HIGHER	or	90 OR HIGHER
HYPERTENSIVE CRISIS (consult your doctor immediately)	HIGHER THAN 180	and/or	HIGHER THAN 120

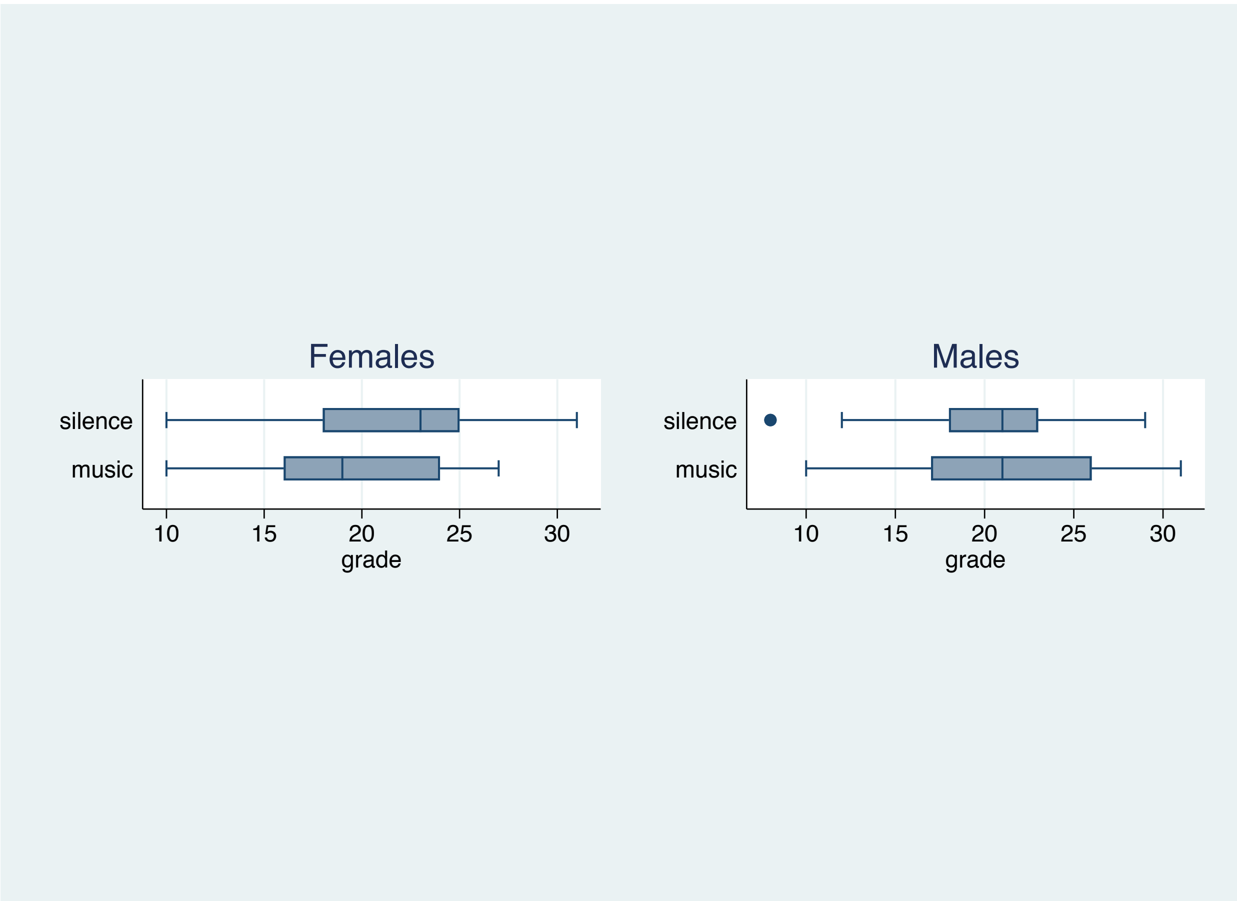
heart.org/bplevels

Figure 2: Mean arterial pressure (MAP) variation by treatment group and gender



Note: The boxplots refer to the variation (in mm Hg) of the Mean Arterial Pressure before/after each treatment.
The dots are outliers and the vertical bar in each boxplot corresponds to the median.

Figure 3: Grade distribution by treatment group and gender



Note: The boxplots show the grades obtained after each treatment.
The dot is an outlier and the vertical bar in each boxplot corresponds to the median.

Table 1: Exam session of Public Economics and participants

Total Number of Students

	(1)	(2)	(3)	(4)
academic year	***1st year Law***	***June Session***	***Exam design***	***% Pass***
t= experiment	233	186	test	0.78
t-1	265	142	open	0.59
t-2	324	146	open	0.58
t-3	391	144	test	0.76

Col.1 total number of first year students enrolled in Law degree.
Col.2 total number of participants at the first summer exam session in June.
Col.3 exam design: 15 multiple choice questions or 3 open questions. Length 1 hour.
Col.4 percentage of students that passed the exam.

Table 2: Descriptive statistics

Variable	**Mean**	**Std. Dev.**	**Min.**	**Max.**
female	0.706	0.456	0	1
age	20.815	2.671	19	42
residence	0.211	0.409	0	1
lyceum	0.701	0.459	0	1
scholarship eligibility	0.244	0.431	0	1
HS grade	82.788	11.644	60	100
grade	20.896	4.941	8	31
cramming	0.532	0.500	0	1
mother working	0.494	0.501	0	1
mother He degree	0.228	0.420	0	1
father working	0.934	0.247	0	1
father He degree	0.217	0.413	0	1
smoke	0.315	0.465	0	1
family illness	0.581	0.494	0	1
listening music	0.451	0.498	0	1
going to concert	0.576	0.495	0	1
using social network	0.456	0.499	0	1

Full sample of 184 students.

Table 3: Balance check of students characteristics across comparison groups

	Silence	Music	diff	Joint P-value
female	0.709	0.703	0.006	0.924
male	0.290	0.296	-0.006	0.924
age	20.784	20.846	-0.061	0.877
residence	0.236	0.186	0.049	0.411
lyceum	0.688	0.714	-0.026	0.700
scholarship eligibility	0.258	0.230	0.027	0.668
HS grade	82.698	82.879	-0.180	0.916
cramming	0.537	0.527	0.010	0.890
mother working	0.440	0.549	-0.108	0.142
mother He degree	0.215	0.241	-0.026	0.668
father working	0.956	0.912	0.044	0.219
father He degree	0.247	0.186	0.060	0.322
smoke	0.311	0.318	-0.006	0.920
family illness	0.569	0.593	-0.023	0.748
listening music	0.430	0.472	-0.042	0.565
going to concert	0.623	0.527	0.096	0.188
using social network	0.451	0.461	-0.009	0.893
exam grade	21.548	20.230	1.317	0.070

Sample of 184 students joining the experiment.

Table 4: Biomarkers values before and after treatment (silence/music)

	Before	Min	Max	After	Min	Max	N
Panel A: All students							
Systolic							
mean	121.005	90	155	118.201	80	170	184
s.d	(12.017)			(13.595)			
Diastolic							
mean	74.532	50	107	72.983	50	105	184
s.d	(10.724)			(10.602)			
Heartbeat							
mean	81.592	50	144	80.440	52	140	184
s.d	(14.066)			(12.181)			
MAP							
mean	92.078	61.121	119.325	89.998	65.988	125.720	184
s.d	(10.421)			(10.654)			
Panel B: Females							
Systolic							
mean	120.238	90	155	117.423	80	170	130
s.d	(11.902)			(13.392)			
Diastolic							
mean	74.723	50	170	72.592	50	100	130
s.d	(10.604)			(10.340)			
Heartbeat							
mean	82.215	50	144	80.592	52	140	130
s.d	(14.909)			(11.819)			
MAP							
mean	91.950	61.121	119.325	89.508	65.988	125.720	130
s.d	(10.459)			(10.540)			
Panel C: Males							
Systolic							
mean	122.851	100	150	120.074	100	150	54
s.d	(12.201)			(14.022)			
Diastolic							
mean	74.074	50	100	73.925	50	105	54
s.d	(11.092)			(11.251)			
Heartbeat							
mean	80.092	56	110	80.074	60	120	54
s.d	(11.784)			(13.120)			
MAP							
mean	92.386	69.937	115.285	91.176	74.265	122.788	54
s.d	(10.4222)			(10.934)			

All values are reported in mmHg unit of measure.

Table 5: Biomarkers variation by treatment group and gender

	Silence	**Music**
Panel A: All students		
Systolic		
% variation	-2	-1
MAP		
% variation	-2	-1
N_{tot} (184)	93	91
Panel B: Subsample Females		
Systolic		
% variation	-3	-0.6
MAP		
% variation	-3	-0.9
N_{fem} (130)	66	64
Panel C: Subsample Males		
Systolic		
% variation	-0.5	-3
MAP		
% variation	0	-1
N_{mal} (54)	27	27

Panel D: MAP decrease

	Female	**Males**
ΔMAP^{-}	81	33
N_{tot} (184)	130	54

N_{tot}: full sample of students.
N_{fem}: total number of females.
N_{mal}: total number of males.

Table 6: Effect of a reduction of test anxiety (MAP variation) on test scores

Dependent variable:	Log Test Score					
	(1)	(2)	(3)	(4)	(5)	(6)
		Females			***Males***	
Δ MAP^{-}	0.136**	0.113*	0.106*	-0.041	-0.047	-0.053
	(0.056)	(0.059)	(0.057)	(0.083)	(0.083)	(0.083)
age		-0.016	-0.018		-0.002	-0.012
		(0.011)	(0.012)		(0.024)	(0.027)
scholarship		-0.086	-0.109		-0.117	-0.134
		(0.068)	(0.069)		(0.130)	(0.138)
father HE degree		-0.079	-0.110		-0.063	-0.061
		(0.075)	(0.078)		(0.104)	(0.111)
cramming			-0.064			-0.100
			(0.058)			(0.103)
listening music			0.044			-0.068
			(0.061)			(0.090)
going to concert			-0.085			-0.015
			(0.055)			(0.084)
using social network			-0.006			0.011
			(0.062)			(0.096)
constant	2.910***	3.309***	3.427***	3.025***	3.114***	3.396***
	(0.046)	(0.242)	(0.269)	(0.067)	(0.491)	(0.583)
N	97	97	97	51	51	51

Robust Standard errors.
Significance levels: *** 1% ** 5% * 10%.
Note: in this model all students are exposed to one of the 2 treatments (silence/music).

Table 7: Effect of the treatments (silence/music) on the reduction of test anxiety (MAP variation)

Dependent variable:		Δ MAP^{-}				
	(1)	(2)	(3)	(4)	(5)	(6)
		Females			***Males***	
silent group	0.186*	0.198**	0.188*	-0.167	-0.182	-0.150
	(0.099)	(0.098)	(0.100)	(0.139)	(0.143)	(0.154)
age		-0.021	-0.021		0.054	0.041
		(0.019)	(0.019)		(0.040)	(0.043)
scholarship		-0.245**	-0.274**		-0.021	-0.056
		(0.121)	(0.123)		(0.177)	(0.172)
father HE degree		-0.232*	-0.238*		0.028	-0.028
		(0.126)	(0.137)		(0.155)	(0.159)
cramming			-0.040			-0.118
			(0.105)			(0.169)
listening music			0.090			0.091
			(0.123)			(0.160)
going to concert			-0.044			0.074
			(0.120)			(0.158)
using social network			-0.163			-0.201
			(0.114)			(0.142)
constant	0.510***	1.070**	1.153**	0.667***	-0.425	-0.070
	(0.071)	(0.425)	(0.443)	(0.093)	(0.836)	(0.980)
N	97	97	97	51	51	51

Robust Standard errors.
Significance levels: *** 1% ** 5% * 10%.

Table 8: Effect of a reduction of test anxiety (MAP variation) on females test scores by treatment group

Dependent variable:	Log Test Score					
	(1)	(2)	(3)	(4)	(5)	(6)
		Silence			***Music***	
Δ MAP^{-}	0.175*	0.178*	0.173*	0.073	0.052	0.029
	(0.102)	(0.097)	(0.089)	(0.074)	(0.081)	(0.083)
age		-0.040***	-0.048***		0.002	0.003
		(0.007)	(0.008)		(0.017)	(0.015)
scholarship		-0.156*	-0.168		-0.027	-0.064
		(0.091)	(0.102)		(0.094)	(0.096)
father HE degree		-0.171*	-0.217**		-0.084	-0.128
		(0.086)	(0.085)		(0.145)	(0.154)
cramming			-0.091			0.003
			(0.077)			(0.085)
listening music			-0.023			0.144*
			(0.083)			(0.082)
going to concert			-0.108			-0.135*
			(0.072)			(0.076)
using social network			-0.010			-0.032
			(0.081)			(0.096)
constant	2.933***	3.854***	4.161***	2.896***	2.884***	2.908***
	(0.096)	(0.196)	(0.219)	(0.048)	(0.365)	(0.348)
N	46	46	46	51	51	51

Robust Standard errors.
Significance levels: *** 1% ** 5% * 10%.

Table 9: Effect of a reduction test anxiety (Systolic blood pressure variation) on test scores

Dependent variable:	Log Test Score					
	(1)	(2)	(3)	(4)	(5)	(6)
		Females			***Females in silent group***	
Δ systolic^{-}	0.165***	0.163***	0.164***	0.169*	0.177**	0.166**
	(0.054)	(0.053)	(0.054)	(0.090)	(0.082)	(0.076)
age		-0.022**	-0.023**		-0.041***	-0.045***
		(0.009)	(0.010)		(0.008)	(0.008)
scholarship		-0.090	-0.111*		-0.184**	-0.204**
		(0.063)	(0.065)		(0.087)	(0.098)
father HE degree		-0.073	-0.100		-0.135	-0.176**
		(0.075)	(0.078)		(0.084)	(0.084)
cramming			-0.075			-0.109
			(0.058)			(0.079)
listening music			0.046			-0.008
			(0.059)			(0.083)
going to concert			-0.074			-0.081
			(0.055)			(0.068)
using social network			0.018			0.027
			(0.063)			(0.077)
constant	2.905***	3.405***	3.486***	2.949***	3.883***	4.098***
	(0.043)	(0.200)	(0.234)	(0.083)	(0.198)	(0.216)
N	97	97	97	46	46	46

Robust Standard errors.
Significance levels: *** 1% ** 5% * 10%.
Note: in models 1-3 females are exposed either to silence or music.
In models 4-6, females are only exposed to silence.

Table 10: Effect of silence on test scores - Reduced Form model

Dependent variable:	Log Test Score					
	(1)	(2)	(3)	(4)	(5)	(6)
		Females			***Males***	
silent group	0.122**	0.129**	0.136**	-0.051	-0.037	-0.038
	(0.054)	(0.055)	(0.056)	(0.080)	(0.084)	(0.081)
age		-0.019*	-0.021*		-0.002	-0.011
		(0.011)	(0.012)		(0.022)	(0.025)
scholarship		-0.110*	-0.133**		-0.101	-0.117
		(0.062)	(0.064)		(0.135)	(0.143)
father HE degree		-0.122	-0.158*		-0.067	-0.064
		(0.077)	(0.081)		(0.102)	(0.107)
cramming			-0.061			-0.087
			(0.057)			(0.103)
listening music			0.053			-0.080
			(0.059)			(0.089)
going to concert			-0.106**			-0.010
			(0.050)			(0.082)
using social network			-0.016			0.026
			(0.063)			(0.096)
constant	2.933***	3.389***	3.521***	3.025***	3.103***	3.350***
	(0.037)	(0.234)	(0.263)	(0.051)	(0.464)	(0.558)
N	97	97	97	51	51	51

Robust Standard errors.
Significance levels: *** 1% ** 5% * 10%.

Table 11: IPWRA model of the effect of test anxiety (MAP variation) on test scores

Panel 1 - Average Treatment Effect		
ΔMAP^{-}		0.121**
		(0.061)
Panel 2 - Outcome models by MAP variation		
	ΔMAP^{+}	ΔMAP^{-}
silent group	0.037	0.139**
	(0.104)	(0.065)
Panel 3 - Determinant of ΔMAP^{-}		
silent group		0.787*
		(0.039)

Females only n. 97.

Panel 1: average treatment effect of a MAP reduction.

Panel 2: regression adjusted models of test scores.

Panel 3: propensity score model of MAP variation.

Table A1: Total number of student at national level

Panel A: Total Number of Students at national level					
	(1)	(2)	(3)	(4)	(5)
academic year	***1st year Law***	***Male number***	***Male %***	***Female number***	***Female %***
t= experiment	17930	6803	37,94	11127	62,05
t-1	19222	7230	37,61	11992	62,38
t-2	22026	8434	38,29	13592	61,71
t-3	23610	8821	37,36	14789	62,63

Panel B: Total Number of Students at MGU					
academic year	***1st year Law***	***Male number***	***Male %***	***Female number***	***Female %***
t= experiment	233	88	37,76	145	62,23
t-1	265	100	37,73	165	62,26
t-2	324	128	39,50	196	60,49
t-3	391	139	35,54	252	64,45

Panel C: Past performance June session at MGU		
academic year	***Female mean score***	***Male mean score***
t= experiment	23.05	22.93
t-1	23.94	21.43
t-2	21.57	19.95
t-3	23.36	22.32

Our elaboration on USTAT-MUR data.