LOSS AVERSION AND GUESSING BEHAVIOUR IN MULTIPLE-CHOICE TESTS: AN EXPERIMENTAL APPROACH

By

Sabuhi Khalili

Submitted to

CENTRAL EUROPEAN UNIVERSITY

THE DEPARTMENT OF ECONOMICS AND BUSINESS

In partial fulfilment of the requirements for the degree of Masters of Arts in Economics

Supervisor: Professor Marc Kaufmann

Budapest, Hungary

2020

ABSTRACT

This paper analyses how individuals with different levels of loss aversion react to the point deduction in multiple choice tests using an online experiment. I compare a penalty system in which 0.5 points are deducted for each wrong answer and participants may skip a question to get 0 score to a baseline system where no such penalty is given, and all other conditions are the same. Participants with high loss aversion levels are expected to skip more questions under the penalty system due to higher expected disutility from point deduction than those with lower loss aversion level. I find that participants randomized into the penalty system skip more questions; however, loss aversion or gender do not play a significant role in determining individual guessing behaviour. Based on the individual interviews with participants, I conclude that lack of high-stake environment in the knowledge competition unlike university or job recruitment exams might have encouraged participants to show similar riskiness on average.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

CHAPTER 1: INTRODUCTION

The need for a quick assessment of large cohorts has made the multiple-choice question format popular in many countries. The multiple-choice question tests in these countries are used in university entrance exams or for recruitment to government agencies as it significantly decreases the time spent for evaluation of answer sheets and leaves no room for the doubts over the objectivity of evaluation.

Along with some benefits, this format has certain drawbacks depending on the conditions it is used in. In practice, multiple choice questions are presented with or without a penalty for selecting a wrong answer. When there is no penalty for wrong answers, test takers are more likely to answer questions, including those they have no knowledge of. Some authorities penalise for wrong answers to decrease the level of guessing and make the results more precise as an estimate of knowledge. In the cases of Azerbaijan (SECRA, 2018) and Turkey (Akyol et al. 2019), ¼ of the score of a correct answer is deducted from final score for each wrong answer. The purpose of this rule is to make the expected score for answering a question with 5 answer choices equal to 0 when students guess randomly. The hope is that students will be indifferent to choosing between guessing and skipping questions if they have no clue on the correct answer and will only guess if they could omit one of the wrong answer choices using their partial knowledge. Although having a positive expected score after eliminating one of the options should encourage students to guess, students may continue skipping questions.

According to the prospect theory, individuals receive bigger disutility from the loss of a certain payoff than the utility from the gain of the same amount in an absolute value (Kahneman & Tversky 1979), which is linked to their loss aversion. Although ¼ of a score awarded for correct answer as penalty for a wrong answer in Azerbaijan or Turkey guarantees a positive expected score as long as a student could omit at least one question, a loss averse student who could omit a few options could still decide to skip a question to avoid point loss. Karle et al. (2019) argue that loss averse students will be less inclined to take a risk by randomly guessing a question and answer a question only if they are confident about it. Because people with more loss aversion put more emphasis on their losses than those with less loss aversion, the disutility from the losses increases as the loss aversion increases. Students at the extreme levels of loss aversion may not risk unless they know the answer to a question certainly. The higher rate of skipped questions is associated with the lower scores as students had a higher chance to answer some of these questions using their partial knowledge. Females in particular are observed to skip more questions than men under the penalty system as they are regarded more risk or loss averse than males which might explain the performance gap (Akyol et al. 2019, Ramos & Lambating 1996, Karle et al. 2019, Baldiga 2014, Ben-Shakhar & Sinai 1991 and Burns & Keswell 2012).

In this paper, I analyse if loss aversion makes students skip more questions in a multiple-choice test where wrong answers are penalized. Before the test, I elicit individual-level loss aversion via a standard methodology and then randomly assign students of different levels of loss aversion and gender into treatment and control groups.

The primary measure of loss aversion is estimated in two levels, high and low. In addition to the randomised treatment, I use difference in difference approach in my estimation to control for other factors (such as participants' knowledge level) and to increase my power. I use Poisson regression model with number of skipped questions as dependent variable to test my hypothesis. Since skipping more questions when guessing a question yields positive expected score, the performance should also be negatively affected by the loss aversion under the penalty system. I test this hypothesis using multivariate OLS regression model with test scores as my dependent variable.

My sample data includes the results of an online knowledge competition where subjects answer multiple choice questions and receive monetary rewards depending on their score ranking. Each participant completes two sessions, and each session consists of 50 questions. The first session includes no penalty for any group, but in the second session based on the loss aversion level and gender I randomly assign participants to the treatment group where each wrong answer is penalised as much as 1/4th of the score for a correct answer and the control group where no such penalty is given.

The findings of the experiment provide evidence that participants may prefer not to answer all the questions despite the positive expected score and skip the questions even under no penalty system. The regression analysis demonstrates that skipping questions significantly affects the performance as expected from the setup of the experiment. However, loss aversion is found to not affect the guessing behaviour under the penalty system as both high and low loss averse participants may skip question at the similar probabilities. The possible mismatch between the level of the stakes used for the measurement tests of the loss aversion level and those for the knowledge competition might explain the absence of this relationship. Based on the individual responses, it is also found that despite the positive expected score some participants (mostly females) shun randomly guessing the questions which seem very unfamiliar to them, and the same people may guess questions about which they know something although this information do not help them to omit any options.

The structure of the paper is as follows. Chapter 2 discusses previous research on this topic. Chapter 3 explains the methodology and experiment design. Chapter 4 summarises the data. Chapter 5 discusses the results for each hypothesis, and Chapter 6 concludes.

CHAPTER 2: LITERATURE REVIEW

One of starting points behind the discussion on the guessing behaviour is the researchers' interest to find out why the performance gap in exams with multiple choice questions format exists between males and females. Historically males have dominated over females for their results at state entrances exams of Turkey (Akyol et al. 2019) and on SAT tests (mainly in the math section) across the United States (Burton et al. 1988, Elsesser 2016). The difference in the guessing behaviour of males and females hints that the knowledge level might not be the only reason behind this performance gap. In the case of Azerbaijan, a similar trend has been observed until recently, but for the last couple of years females get significantly higher scores than males on average (SECRA, 2018)[1]. The case of Azerbaijan demonstrates that the gender gap might be diminishing for some cohorts or a gap where females are doing better may exist. It is not likely that women have undergone a character change with respect to their guessing behaviour, there might be other reasons behind this phenomenon. Azerbaijan for instance has a growing female population that shows greater interest in education and career which explains why they have closed the gap in the exam results, but even in this situation the penalty system might have hindered their otherwise much higher results.

In addition to the observed data, field experiments (including those by Ramos & Lambating (1996), Karle et al. (2019), Baldiga (2014), Ben-Shakhar & Sinai (1991) and Burns & Keswell (2012)) demonstrate that females skip significantly more questions under a penalty system and this in turn negatively affects their results. They also provide evidence for women being more risk or loss averse than men on average.

[1] For year 2019, the average score of females is 11 percent higher than males. The acceptance ratio and number of accepted people are very close for two groups nevertheless (this is due to the fact that slightly less than half of the male applicants get a score between 0-100 on the scale of 700)

The researchers are not interested in the skipped questions if these are the questions students have no clue on. In such questions, random guessing yields zero expected score in university entrance exams and whether a student skips them or not should not matter much on aggregate. However, a student's performance will be negatively affected if some underlying factors discourage a student from guessing a question about which he knows enough to omit a few options. To test if a student would still skip a question with a positive expected score, most field experiments use exam settings with three or four answer options instead of five for each question. This establishes an artificial case of all students having enough partial knowledge about the question to omit one option. Researchers later check if students preferring to skip rather than attempt a question under these settings have any differences from those who do not. The variations in loss and risk aversion and confidence levels are suggested in the literature below among the differences that could play an important role on determining student's guessing behaviour.

Akyol et al. (2019) use a structural model for university exam results in Turkey and demonstrate that the penalty for selecting wrong answers in the exams with multiple choice questions reduces the expected score of test takers with higher risk aversion as they continue to skip questions despite their partial knowledge.

Karle et al. (2019) use a microeconomics exam for a field experiment with the participation of around 600 students and find that women skip more questions under a penalty system. The authors show that the performance of students is negatively affected due to different guessing behaviour, which could partially explain the performance gap. The loss aversion, but not risk aversion, is found to significantly affect guessing behaviour, and hence a student's performance. The authors use mixed lotteries for measurement of loss aversion where participants may end up losing as a result of their lottery choices, and comparisons of lottery choices with only gains is utilised for the measurement of risk aversion. Since the penalty system involves losses in case of wrong answers, loss aversion seems a better fit and this might be one of the reasons of why they find no significant relation for risk aversion. However, the authors do not use risk and loss aversion in the same regression due to technical issues and end up running separate regression for them with a few strong assumptions. The overall validity of the results depends on how one accepts these assumptions. In addition, Karle et al. (2019) suffers from possible omitted variable bias for relying on an observational data for checking performance differences where unobservable factors such as ability or previous knowledge on the related subject (math for introductory economics course) could be correlated with the test outcome. The authors use a cognitive reflection test to handle this shortcoming, but this only covers the problem partially as this is a general test on cognitive skills which cannot fully capture a student's ability or knowledge level for a given subject.

Baldiga (2013) conducts a field experiment with 406 participants where they answer SAT questions under a low, high and no penalty system. The findings of this paper demonstrate that men guess more than women in the penalty treatment. Knowledge, risk aversion and confidence are presented as the factors partially explaining this behaviour. However, there are a few shortcomings of this experiment. Participants are not randomised to different treatments within a session, and at each session different set of people answer the questions. Although similar questions across the sessions helps with the comparison of different groups, the participation of different individuals at each session and non-existence of a randomisation process at the treatment level raise questions about the exogeneity of the explanatory variables. The measurement of knowledge as one of the control variables on the other hand is based on a participant's correct answers to the same set of questions at a later stage of the same session where answering questions are mandatory. This implies that participants had to guess the questions they didn't know for sure. As a result, the measurement of knowledge resembles a typical measurement error, where the estimated knowledge is the correct answers given by participants based on their knowledge plus random error, i.e. randomly given correct answers through guessing. In the pilot sessions of this experiment, the participants are given a chance to decline answering by choosing an option “I don’t know”, but for some reason, the researcher decided to omit that option in the main sessions.

Ben-Shakhar & Sinai (1991) and Krawczyk (2011) check if the way the treatment of penalty (or no penalty) is presented makes any difference in students’ test results. Ben-Shakhar & Sinai (1991) conducts the field experiment among 604 children in Israeli schools and 300 university applicants. Both groups complete tests with multiple choice questions with no punishment, but only the second group is explicitly instructed that there is no penalty for a wrong answer if a student tries to guess. The results show that females guess less than males regardless of subject or instruction/treatment type. The males also perform better than females in most subjects and guessing partially explains the performance gap. Krawczyk (2011) uses a Microeconomics exam at a university in Poland to conduct his field experiment to test if the willingness to guess is affected by the treatment type of exam. Two types of treatment are implemented: Loss Treatment, where students are deducted a point from initial score based on their responses (deduction of 0 for correct, -2 for skipped, -3 for wrong answer) and Gain Treatment, where students receive points based on their answers (additional 3 points for correct answer, 1 point for skipped question and 0 point for wrong answer). The results show that the type of treatment does not significantly affect the guessing behaviour, but females skip more question than men. This also implies that loss-aversion is not an important factor, but some other condition specific to women may have a better explanation for their guessing behaviour. The findings of these two papers also show that participants, especially women, may prefer to skip questions under different exam settings including in a no penalty system. Since there is no added cost for answering question, skipping a question at this level may sound as more irrational than to skip a question with a positive expected score under a penalty system. This suggests that forces behind guessing behaviour may continue influencing individual decision under any exam settings and have a stronger impact through the penalty system.

Niederle & Vesterlund (2007) claim that the difference in confidence levels is the main factor behind why men take more risks than women. The authors form groups of two men and women in an experiment to solve a simple task of summation of two-digit numbers. At one of the experiment's stages, they are given a chance to choose whether they want a guaranteed small payment in a single player setting or get 4 times as much as the previous amount by choosing tournament setting and becoming a winner. Under these settings, the expected earnings from joining the tournament is the same with the guaranteed earning under single player setting (this very much resembles a decision-making problem of guessing in multiple choice questions). The results show that 73 % of men decide to join the competition while only 35 % of women make the same decision. The researchers find that overconfidence among men is the primary reason for this gap, while risk and feedback aversion have a little role in it and no relation of the performance. The authors however do not necessarily use a proxy for risk aversion in their models, rather use the results of confidence as a counter argument against risk aversion. It is very likely that risk aversion or knowledge is correlated with overconfidence, and the significant coefficient of confidence might then be biased due to the omission of these variables. The findings of this paper contradict Baldiga (2014) which presents all these variables as factors behind the guessing behaviour.

Among all the factors given in the literature, my primary focus in this paper is to analyse whether loss aversion impacts performance through the differences in their guessing behaviour.

If men indeed like taking more risks than women as stated in the literature above, the riskiness then should account for some of the performance gap between these groups. If this claim was not true, but it is also found out that risk preferences indeed affects the decision to guess a question or not, the nature of issue this time would shift from gender level to the discrimination of a certain group (group of people with more conservative risk preferences for instance) which is hard to explicitly identify. Approaching the problem from a gender perspective on the other hand makes the problem more salient and clearer for majority of people dealing with this at the administration level. I also take the gender side of the issue into account and design my experiment accordingly. Among risk preferences I focus on the loss aversion level of individuals as the guessing in a penalty system resembles lottery choices with losses more.

CHAPTER 3: METHODOLOGY AND EXPERIMENT DESIGN

3.1. THEORETICAL FRAMEWORK

A student taking an exam with multiple choice questions will answer a question if the following inequality is satisfied:

$$p * U(x) - (1-p) * \lambda * U(x * d) \geq 0$$

$$where\ p = 1/number\ of\ remaining\ options\ and\ \lambda \geq 1$$

Here x represents the score received by answering a question correctly, d is the amount of deducted point for a wrong answer as a fraction of the awarded score for a correct answer. Probability p of answering a question correctly decreases as the number of answer options increases. I assume if test takers do not know the answer, they will randomly guess among the remaining options after omitting the incorrect ones with their knowledge. λ multiplies the size of any disutility received from the point losses due to a wrong answer making it bigger than the utility received from the gain at the same amount. For loss averse individuals, λ is strictly greater than 1 and it increases as the individual's loss aversion level increases. For individuals who are not loss averse, λ is equal to 1, i.e. the utility received from a certain amount is equal to disutility received from the same amount. By setting lower level of λ equal to 1, I assume that there is no individual who loves making losses.

The inequality above implies that loss averse students will not guess a question if there is a penalty for guessing it wrong and λ is big enough to make the overall utility negative. A student who knows answer to a question certainly will have only one remaining option, which is the correct answer, and λ will be ineffective as p is equal to 1. On the other hand, loss aversion level may be so high for some loss averse individuals, they may decide to skip a question when the number of remaining options is only two, which yields a high positive expected score in most penalty systems in practice. And if a test taker with a high loss aversion skips significant amount of questions despite the positive expected score, then he will get a lower score than another student with the same level of knowledge who skips less questions due to lower loss aversion level. The following two hypotheses summarise the arguments above about the impact of loss aversion on the guessing behaviour and the performance under the penalty system:

Hypothesis I: When there is penalty for wrong answers, loss averse test takers will answer less questions in total.

Hypothesis II: Since those with higher loss aversion skip more questions, in the presence of penalty they will have lower score than those with lower loss-aversion

The penalty system would affect females more if there were more loss averse females than males. It is also interesting to see how other gender related factors may affect guessing behaviour and performance under the penalty system if this was not true. The next two hypotheses summarise my predictions for the gender related differences:

Hypothesis III: If there are the same number of loss averse people among males and females, females will not skip more questions than males.

Hypothesis IV: If there are the same number of loss averse people among males and females and women do not skip significantly more questions than males, the penalty system will not have negative impact on performance of females more than males.

3.2. EXPERIMENT DESIGN

My initial plan was to conduct a lab experiment at Central European University, but due to the COVID-19 outbreak at the time of writing this paper, I had to move everything to an online knowledge competition format. The students had to fill in the application form to participate in the event and then wait for the event time to join the competition through their own devices. The application form included demographic questions and a mixed lottery choice question (similar to Karle et al. 2019, see Appendix A) which would define their loss aversion level. There were 4 loss aversion levels (later grouped as high and low) the participants could fit into based on their answers. The number of loss aversion level indicates the number of accepted lottery choices in the survey where more accepted lotteries imply smaller loss aversion level. Before the event started, I matched each participant with another participant with the same loss aversion level and gender and randomly assigned one of them to the treatment group and the other to the control group.

The competition/experiment was conducted in two consecutive sessions. At the beginning of each session, each participant received a popup message on the game app about the scoring system of the session. In the first session, all participants answered 50 general knowledge questions with 4 answer options without a penalty for a wrong answer and each correct answer was awarded with 2 points. In the second session, the participants still had to answer 50 questions, but this time the treatment group received a different popup indicating that they would be deducted 0.5 points for each of the wrong answers, while a correct answer was still worth 2 points. The scoring system for the control group was the same with the first session. All participants were ranked from the highest score to the lowest and they received an award for each session separately depending on their ranking[2].

Since each question has four options, regardless of the presence of the penalty system or how many options a participant could omit, the expected score from randomly guessing a question

[2] Participants at the Top 51%-100 % received per session 500 HUF, Top 26% -50% 875 HUF, Top 6%-25% 1,125 HUF and Top 5% 1,500 HUF. Minimum award one can win from the competition was 1000 HUF and the maximum amount was 3000 HUF

remains positive through the entire game. This requires participants to answer all the questions if they only care about maximising their performance.

3.3. REGRESSION MODELS

My regression analysis is divided into two main parts in line with my hypotheses: the impact of loss aversion on guessing behaviour and performance under the penalty system (Hypotheses 1 and 2), and the impact of gender specific differences on guessing behaviour and performance under the penalty system (Hypotheses 3 and 4).

The first part includes these two models below for Hypotheses 1 and 2 accordingly:

$$\textit{Skipped questions}_i = e^{\beta_0 + \beta_1 \textit{penalty}_i + \beta_2 \textit{loss-averse}_i \times \textit{penalty}_i + \beta_3 \textit{loss-averse}_i + \beta_4 \textit{second stage dummy}_i + \epsilon_i} \quad (1)$$

The dependent variable in model (1) is guessing behaviour and I use s*kipped questions* (the number of questions skipped in total by an individual in a session) to proxy for (the opposite of) guessing behaviour. The theoretical limit of skipped questions ranges from 0 to total number of questions available in a test. Since the participants have an incentive to get higher in the ranking and skipping most questions guarantees defeat, most participants should end up at lower levels of number of skipped questions and the amount of people at the higher levels of skipped questions should gradually decrease towards total number of questions. Such distribution of number of skipped questions requires me to use count outcome models such as Poisson regression model for Model (1).

Loss averse refers to a dummy variable which is equal to 1 for individuals at high levels of loss aversion who accepted 3 or 4 lotteries. β_2 demonstrates the impact of penalty system through loss aversion and it is expected to have a positive sign. This implies that loss averse people skip more questions than other individuals under penalty system. β_1 represents the remaining impact of the penalty system on guessing behaviour. Since not loss averse people in the definition of loss averse dummy variable still has a low, yet some loss aversion level, this coefficient might have positive sign. If the loss aversion level of not loss averse individuals in my sample is so low that the expected utility from randomly guessing a fully unknown questions is still positive, then this coefficient should be zero. β_4 captures any aggregate level change that happened between two sessions other than introduction of the penalty system. Since the competition stages are conducted right after one another, the only thing that may have changed is the difficulty of questions. Although questions are randomly selected from the same database, there is a chance that questions might be harder in one of the sessions than the other. Later I separate my sample to men and women to compare how these estimates from equation (1) differ based on gender. Loss aversion should not have a separate effect on women than men.

The below is given the second model of the first part where I change the dependent variable in the Model (1) to test Hypothesis 2.

$$Performance_i = \beta_0 + \beta_1 penalty_i + \beta_2 loss\text{-}averse_i \times penalty_i + \beta_3 loss\text{-}averse_i + \beta_4 second\ stage\ dummy_i + \epsilon_i \qquad (2)$$

I use a multivariate linear regression model since the performance should have a normal distribution as the number of observations increase. The participant's test score is used as proxy for performance which has theoretical range between 0 and 100. β_2 should have negative sign if it is found in Model (1) that loss averse participants skip significantly more questions under the penalty system. β_1 represents the impact of the penalty system on test scores through the point deduction plus potential score loss for skipping more questions due to any other factor other than loss aversion. This coefficient should have a negative sign, too. If there is any difference between the difficulty of questions or any other change in environment other than the random assignment of the penalty system, β_4 should capture it as in Model (1).

Since the participants of the experiment have an equal number of people for each gender and loss aversion level, the presence of any other effect due to gender might be captured by these two models below.

$$\textit{Skipped questions}_i = e^{e\ \beta_0+\beta_1 penalty_i+\beta_2 female_i \times penalty_i+\beta_3 female_i+\beta_4 second\ stage\ dummy_i+\epsilon_i} \quad (3)$$

I use Model (3) to test Hypothesis 3, in which I use number of skipped questions as dependent variable. I run Poisson regression model as it is in Model (1).

β_2 represents the difference between number of skipped questions between men and women due to the penalty system. Since I have balanced number of loss averse people among males and females in my sample, β_2 should capture any other impact of penalty system on skipping behaviour due to the gender differences. The coefficient may be positive if such differences makes the women skip more under the penalty system. β_3 captures any impact of gender on guessing behaviour that is not specific to the penalty system. If the knowledge level of females is higher than males on average, β_3 will be pushed downwards as those with higher knowledge skip less questions, but the knowledge level is only one of the factors that may not be evenly distributed between males and females. β_1 captures the impact of penalty system at individual level including the impact of loss aversion. The sign may take any form depending on whether these individual level factors encourage or discourage guessing behaviour in aggregate. However, I expect this to be positive as the factors such as confidence, risk aversion, knowledge and loss aversion should encourage people to skip more questions. The interpretation of β_4 is the same with the previous models.

$$\textit{Performance}_i = \beta_0+\beta_1 penalty_i+\beta_2 female_i \times penalty_i+\beta_3 female_i+$$

$$\beta_4 second\ stage\ dummy_i+\epsilon_i \qquad (4)$$

Model (4) tests Hypothesis 4 with the same set of independent variables. β_1 will capture the impact of point deduction plus point loss as a result of more skipped questions due to any underlying factor including loss aversion. β_2 will take the opposite sign of the same variable at Model (3). If there is no gender related difference in guessing behaviour under the penalty, this coefficient should be zero. The sign of β_3 changes depending on how the factors affecting guessing behaviour regardless of the penalty system and the knowledge level are distributed among males and females. The interpretation of β_4 is the same with the previous models.

Since I randomly assign the penalty system based on loss aversion and gender, loss averse and female dummy variables in Model 1-4 are not necessarily exogenous on their own, neither they form a main part of my analysis. Interested reader should take the interpretation of the coefficient estimates of these variables in Chapter 5 with caution.

There is one final model I estimate to see if skipped questions has negatively affected the performance. This is a mere technical estimation to demonstrate that skipping more questions is associated with lower score after controlling for the knowledge level. Model (5) below describes this relationship:

$$Performance_i = \beta_0 + \beta_1 treatment_i + \beta_2 skipped\ questions_i + \beta_3 ranking_i + \epsilon_i \quad (5)$$

I use the ranking[3] of the first session to control for the knowledge level, and the treatment dummy is used to control for the losses due to the point deduction. The coefficient of number of skipped questions is expected to have negative sign as guessing yields positive expected score under the experiment settings. If skipping negatively affects performance, then I can set a theoretical foundation for if and why loss-aversion affects performance under the penalty

[3] Ranking is based on the score, if two participants had the same score, I gave a better place in raking to the participant with less skipped questions. This is only done for regression analysis, participants in the competition however is ranked based on their scores alone.

system. Otherwise, there is no other logical explanation how one's loss or risk aversion may interact with the performance under penalty system differently from any other system.

There are a few points specific to the design of the experiment and methodology that I want to discuss in the following sections of this chapter.

3.4. THE MEASUREMENT OF LOSS-AVERSION

In line with Baldiga (2013) and Karle et al. (2019), I use a mixed lottery to determine loss aversion of test takers. There are usually two approaches in using lotteries for measurement of risk preferences. Some researchers (including Schubert, et al. (1999), Moore and Eckel (2003), and Harbaugh, Krause and Vesterlund (2002)) prefer providing lottery choices (usually with high stakes) which do not involve actual monetary transactions. The rest, like Karle et al. (2019), present a case where participants will indeed lose or earn money based on their lottery choices. Due to cost issues, they set the stakes very low. The former case makes it hard to believe that participants will indeed give the same decision when they have been presented with the same choices in real life. The outcome of the latter case on the other hand cannot be safely generalised to scenarios with high stakes without making strong assumptions.

Overall, it is debatable what is the right level of stakes at these lotteries. If I set the stakes too high, my participants' loss aversion levels could converge to a level where they would not risk a sure gain of a prize big enough against any risky choice; if the stakes are too low, they may be indifferent between lottery choices and make their choice randomly. If participants on the other hand are not economics or math students and see such lottery choices for the first time, they probably wouldn't want to spend much time on understanding a complicated lottery choice question which would decide if they win five or seven euros. The situation becomes more problematic if non-homogeneity in income levels is allowed as stakes are different among participants. In this case, the awards allocated in the test may be high for one income group, small for the other, and the experiment designer may not achieve a similar level of competitiveness among all participants.

To avoid high cost and experimenting with low stakes, I set the monetary reward in lottery choices as high as the budget would allow for all participants together and let only one randomly selected participant to play the lottery. This way participants are expected to make their decisions as close to reality as possible without incurring a lot of cost. Since the participants are students, income level should not matter much and a lottery that could win a prize of 20 000 HUF (around 57 euros) should neither be too high nor too low to encourage students to make a careful choice. Through application, I could get participants for all categories and more than 90 % of participants spent more than 3 minutes to fill in the application form and half of them watched the supplementary video material to understand the lottery choices better. Despite this information and all the effort I put into making the lottery choice question as simple as possible, I wouldn't deny the possibility that few participants did not understand the lottery question and made a random choice in their answers. This is the probably one of the biggest problems faced by all the researchers trying to measure risk or loss aversion through survey.

Baldiga (2014) on the other hand follows a different approach and ties the lotteries to the test scores of participants. In her design of the lottery, participants decide accept a lottery which earn for them an additional point for the question if the randomly drawn number is not bigger than target number for that specific lottery; otherwise they lose 0 points if they are in control group or ¼ point if they are in treatment group. This approach may create a conflict between measurements of explanatory variable (risk preferences) and dependent variables (number of skipped questions). If a student were to gamble a lottery for getting another point, she would do the same by just taking the same gamble by answering question which has positive expected score. As pointed out by the author, accepting a lottery with a target number of 75 is equivalent to deciding to guess a question with 75% probability of getting it right. Either for this particular issue or for some other problem envisioned by the author, the questions with lottery choices are provided in the second part of the experiment, so that the questions with and without the lottery choices are not the same. If using different questions sets was a valid approach to handle the measurement issue, the author could have just used the number of skipped questions in one of the parts of the experiment for measurement of risk preferences and the other for skipping behaviour. I believe the measurement of risk and loss aversion would better be identified in monetary settings.

3.5. INCENTIVES FOR COMPETITIVENESS

Although I could increase the stakes for loss-aversion test, I couldn’t do the same thing for the knowledge competition due to the budget issues. Instead of using the whole budget to pay the rewards for top 5 percent, I paid a minimum prize to everyone who joined the event plus an additional reward based on their ranking. This made me allocate less money to top scorers making it a low stake competition. In contrast to the loss aversion test, I expected that low stakes of competition should not affect results significantly as those who already spend time on it would like to be higher in the rank for their pride at the very least or would reflexively copy their usual responses to penalty system at exams with more pressure. However, I realised after the experiment that this did not work for all participants and I explain my thoughts on this at Chapter 5.4 and 6.

I could instead increase the prize money by limiting the whole prize pool to top five percent to increase the stakes of the competition, but this time those who realise that they don’t know most of the questions would lose the hope and quit participating during the event which is a serious problem to handle. As a result, at the cost of some competitiveness I ensured that nobody drops out for such reasons.

3.6. CHEATING AND ITS IMPACT ON THE RESULTS

Since the experiment is held online, the participants could find a way to cheat. It was almost impossible for a user to hack the online system and change their results, and the 20 second limit for answering each question made it very difficult to use Google search for finding answers. However, as for many other online events there was no remedy against multiple people using a single account to answer questions. Under certain conditions, this wouldn't necessarily alter the results. If the participant deciding whether to answer or skip a question is the same person who filled in the application form (which is very likely scenario), any external help would only increase the knowledge level of that participant and not intervene in his guessing behaviour given that there is enough time left to decide to skip or answer the question. Since I am only interested if the penalty system would affect a guessing behaviour and through that alter their results, constant increase in knowledge level doesn't have any importance for this analysis as long as the same external support is provided in both stages. Since participants had to answer the second stage right after they answered the first stage, I expect that if anyone used external support then the same support was available through entire game.

However, there are other likely conditions which could alter my results. In one of these cases, the 20 seconds limit would expire while the participant tries to search the answer in the Internet. My expectation on assigning this limit was that participants would stop cheating after one attempt realising the timeout would not allow cheating at a large scale. One of the participants even complained to me about the timeout for his failed attempts to find the answers in Google and dropped out of the event right at the beginning. After the event was complete, I found another participant who skipped notable amount of questions in both sessions and answered all the remaining questions correct. The participant's change in number of skipped questions from one session to another was not his reaction to the penalty system, rather it was about how much answers he could find (probably) on Internet. Even if I claim that participant's decision to skip the questions he does not know may constitute a valid decision at some level, however, this result is void because my assumption on the same level of external support is clearly violated and increase in number of skipped questions is only because of the change in this level. This was the only case that I had strong evidence for cheating, and I dropped the participant's results.

3.7. THE USE OF GENERAL KNOWLEDGE QUESTIONS

The use of general knowledge questions in the experiment helps me to have almost the same expected score for almost all questions for all participants. These are the questions that a participant would hardly make any guess using their partial knowledge and thus expected score will remain the same for all the questions and be positive only because I use four answer options instead of five.

3.8. DATA CLEANING

I prefer to speak of the data cleaning in this chapter because it concerns my experimental design. In the first version of the game software I used for the knowledge competition, there was confirmation pop-up that would appear after participants select to submit their answer or skip the question. This pop-up practically decreased the amount of time to answer a question from 20 to 15 seconds which is very small for some questions. In the first release of the game in a smaller group, 3 participants had between 1 and 5 questions skipped because of the timeout and I made a random choice for these questions once the participants confirmed that they would answer the question if not for the timeout. I removed the pop-up for the main event. Two more participants claimed they had no skipped questions and I asked them to complete the session again due to slightly larger number of skipped questions. There was no significant increase in scores in both cases. After the main event, I contacted all the participants (sometimes individually if necessary) to see if they for some reason did not read the message about the penalty system or did not know the skip button gives them a chance to not answer a question.

This was all the information I asked, and five participants confirmed that they missed this information. I asked them to do a third session with the same reward possibilities to keep the incentives intact. The questions were also the same although I kept this secret until the session. Some of these participants guessed all the questions as it was in the previous session and some decided to skip.

CHAPTER 4: DATA

4.1. SUMMARY STATISTICS

34 males and 32 females participated in the experiment. Out of 66 participants, 32 participants belong to high (first and second) loss aversion levels labelled as loss averse. Exactly half of the females belong to the loss averse participants; for the males this number is equal to 18. Half of the loss averse members have been assigned to the treatment group through randomisation and the other half is assigned to the control group. There is slight disbalance in the numbers because of one dropped result (the participant was found to be cheating through the entire game).

Table 1-2 describe the summary statistics of participants based on gender and loss aversion level. Table 1 indicates that participants skipped questions in session one (albeit much less than the second session) in which there was no penalty system for any participants. Since the higher number in the loss aversion level indicate lower loss aversion, the number of skipped questions should decrease with decreasing loss aversion. This pattern is not fully observed as the figures demonstrate that the penalty system encourages participants to skip more, however, there is not much difference between different levels of loss aversion.

Table 2 shows that females skipped less questions than males in both sessions, and males dominated the first session based on the score and were slightly outperformed in the second session. I use adjusted score to make the performance results comparable across the treatment and control groups. For my small sample, there is a performance gap only in the first session where the skipping behaviour is similar on the average. In the second session, this performance gap might have been closed down due to the risk loving behaviour of females which suggests skipping questions may have affected the performance, but this time males are the population taking the less risk.

Table 1. Summary Statistics

Session 1	Loss Aversion Level			
	I	II	III	IV
Score	43.29	50.11	54	47.60
	(9.302)	(12.26)	(14.54)	(8.757)
Skipped Questions	0.143	2	1.474	1.733
	(0.535)	(4.102)	(3.717)	(4.949)
Correct Answers	21.71	25.06	27	23.80
	(4.697)	(6.131)	(7.272)	(4.379)
Age	24.79	25.22	24.68	25.80
	(3.043)	(2.439)	(2.829)	(4.663)
Observations	14	18	19	15

Session 2 Control Group	Loss Aversion Level			
	I	II	III	IV
Score	46.29	42.80	49.56	34.75
	(5.469)	(12.44)	(6.307)	(6.135)
Skipped Questions	0	1.600	0.444	0.875
	(0)	(3.204)	(1.333)	(2.475)
Correct Answers	23.14	21.40	24.78	17.38
	(2.734)	(6.222)	(3.153)	(3.068)
Age	23.86	25.60	25.11	26.50
	(2.340)	(2.716)	(2.315)	(5.555)
Observations	7	9	9	8

Session 2 Treatment Group	Loss Aversion Level			
	I	II	III	IV
Score	23.71	34.50	38.05	27.21
	(6.512)	(12.14)	(21.04)	(6.800)
Skipped Questions	16	13.88	13.0	15.14
	(14.50)	(8.967)	(10.34)	(8.533)
Correct Answers	16.29	21.0	22.60	17.86
	(3.352)	(5.904)	(9.582)	(3.237)
Age	25.71	24.75	24.3	25.0
	(3.54)	(2.12)	(3.30)	(3.65)
Observations	7	8	10	7

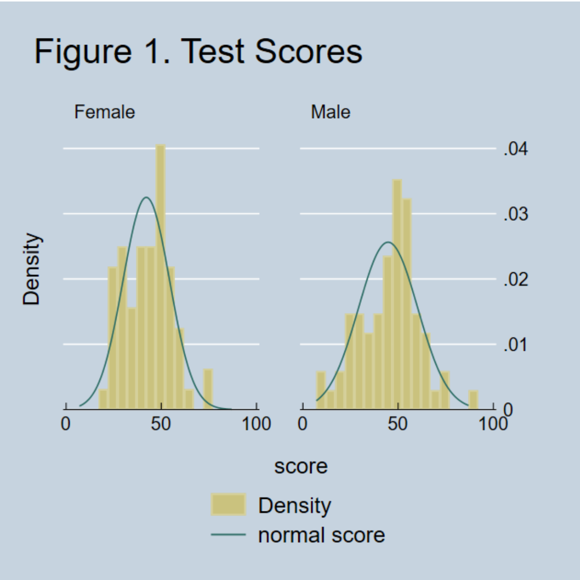
mean coefficients; sd in parentheses; skipped questions and correct answers are measured in integer numbers within 0-50 interval, score is measured with points within 0-100 interval.

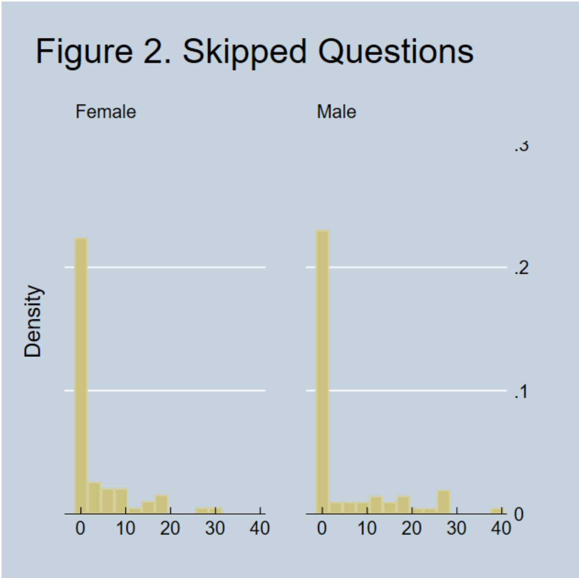
Table 2. Summary Statistics on Gender Allocation

Session 1	Female	Male
Score	46.13	52.18
	(12.14)	(11.51)
Skipped Questions	1.375	1.412
	(3.180)	(4.258)
Correct Answers	23.06	26.09
	(6.069)	(5.754)
Session 2	Female	Male
Adjusted Score[4]	42.13	41
	(9.876)	(13.45)
Skipped Questions	5.938	8.706
	(8.493)	(11.14)
Correct Answers	21.06	20.50
	(4.938)	(6.725)
Observations	32	34

mean coefficients; sd in parentheses

4.2. DENSITIES OF DEPENDENT VARIABLES





The distributions of the proxies of the dependent variables are given in Figure 1-2. Test score is not exactly distributed like a normal distribution (Figure 1), but this is mostly because I have

[4] Adjusted score is measured through recalculating score without point deduction

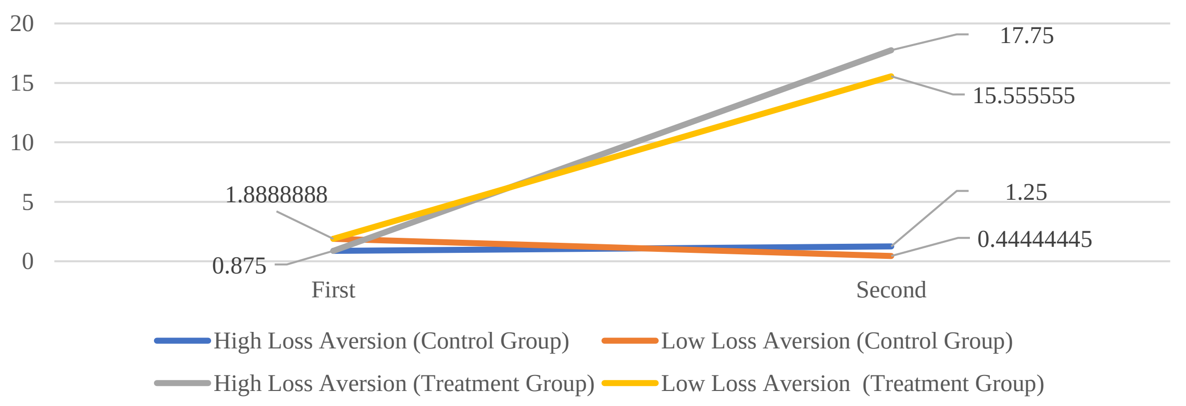
limited sample data. However, the current shape of it hints in favour of normal distribution in large samples.

The distribution of number of skipped questions indeed resembles Poisson distribution which suggests that Poisson regression might be good fit for Model 1 and 3.

4.3. DIFFERENCE IN DIFFERENCE

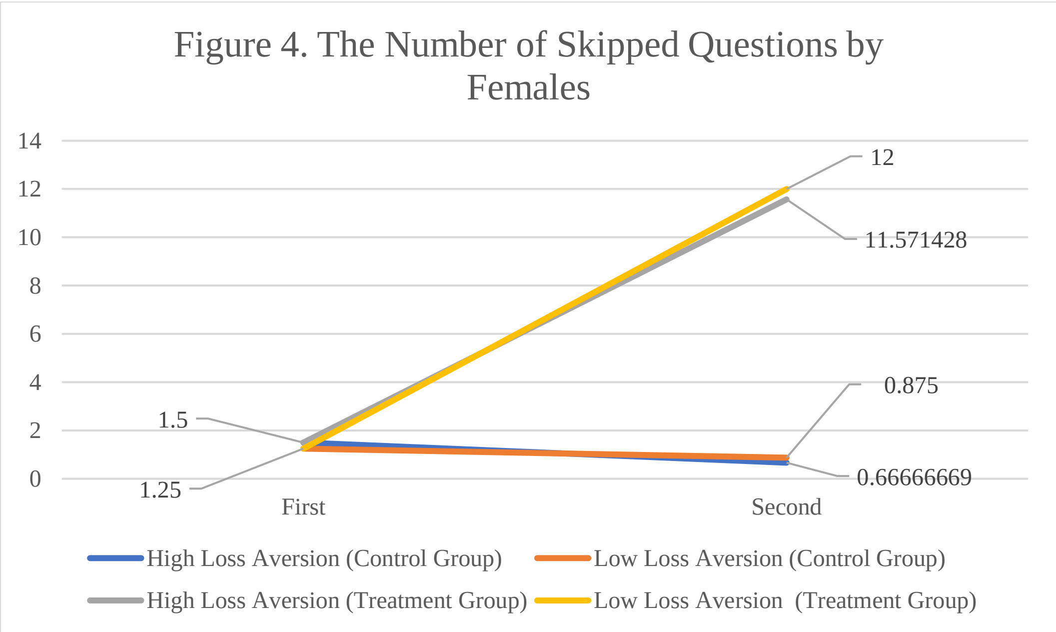
Figure 3-4 present how loss averse males and females reacted to the penalty system in the second session. I have only one figure for treatment and control groups at the first session, because I apply the penalty only in the second session. Figure 3 shows that males reacted to the penalty system and skipped significantly more questions than those in the control group. Although the number of skipped questions by males with high loss aversion level (loss averse in my definition) under the penalty system is bigger than those by males with low loss aversion level, the regression results at Chapter 5 demonstrate that the difference is not significant.

Figure 3. The Number of Skipped Questions by Males (mean)



Females also negatively reacted to the introduction of penalty system and skipped slightly less questions than males in the treatment group, but this difference is also not significant according to regression results at Chapter 5. However, this small difference between female and male participants in the number of skipped questions helped the females to close the performance gap from the session 1(see Table 2).

The gap in number of skipped questions between females with high and low loss aversion levels under the penalty system is even smaller than the gap between the same groups at the male population. This implies that neither male nor female population provide evidence for loss aversion to explain variation in the number of skipped questions under the penalty system in my experiment. The following chapter about the results confirm all these finding through regression models.



CHAPTER 5: RESULTS

5.1. THE IMPACT OF GUESSING BEHAVIOUR ON PERFORMANCE.

Table 3 presents the results of Model (5). The first and second columns presents results with score as proxy for performance, and the last column those with the adjusted score. Since the second column includes only treatment group participants in the sample, *penalty* variable is 1 for all observations. Therefore, I omitted penalty in this model, and this made constant lower due to negative impact of penalty on score through point deduction.

Table 3. The Impact of Guessing on Performance

Dep var: Score	ALL b/se	Treatment Group b/se	All(adj. score) b/se
Skipped Questions	-0.438*	-0.423[5]*	-0.741***
	(0.18)	(0.22)	(0.15)
Ranking	-0.326***	-0.365**	-0.285***
	(0.07)	(0.11)	(0.06)
Penalty	-8.108*		4.197
	(3.51)		(3.03)
constant	55.399***	48.194***	54.179***
	(2.97)	(5.06)	(2.57)
BIC	507.377	259.266	488.136

* p<0.05, ** p<0.01, *** p<0.001; beta coefficients, se in parentheses; the sample includes only session 2 results

The results indicate that an additional skipped question lowers down score by 0.42 points with everything else constant. Since the penalty system's direct impact on the performance is through the point deduction after controlling for skipped questions, penalty variable is only significant when I use score and it becomes insignificant when I use adjusted score. The impact of skipped questions has increased in the last column, because skipping question is more costly when there is no point deduction for wrong answers.

[5] Significant at 0.06 level, considering the small sample size, I consider this coefficient significant as well.

There are two take-aways from these results. First, they confirm that skipping questions has significant and negative impact on the performance. Second, the results are consistent across groups and models (including those below) and there is no illogical pattern I observe. I can now move to the main part of the analysis as I see no systematic issue in my sample with respect to the exogeneity of assignment of the penalty system.

5.2. THE IMPACT OF LOSS AVERSION ON GUESSING BEHAVIOUR AND PERFORMANCE UNDER THE PENALTY SYSTEM (HYPOTHESIS 1-2).

The results of Poisson regression in Table 4 demonstrate that there is no significant relation between loss aversion and guessing behaviour under the penalty system as I cannot reject null hypothesis of no effect. The results are similar across the genders as expected. The positive and significant sign of penalty implies that the penalty system indeed affects the guessing behaviour and the impact is very large. Contrary to the previous findings, the impact of treatment on the penalty system has remained the same across the genders, in fact the coefficients and standard deviations are very close to each other. These two findings combined indicate that penalty system encourages participants to skip more questions despite the positive expected score, however, the impact is not gender specific.

Table 4. The Impact of the Loss Aversion on Guessing under the Penalty System

Dep Var: Skipped Questions	All	Males	Females
	b/se	b/se	b/se
Loss Averse	-0.139	-0.342	0.079
	(0.18)	(0.26)	(0.27)
Loss Averse x Penalty	0.207	0.474	-0.115
	(0.21)	(0.29)	(0.31)
Penalty	2.794***	2.792***	2.796***
	(0.22)	(0.30)	(0.33)
Second Session Dummy	-0.561*	-0.539	-0.589
	(0.22)	(0.30)	(0.32)
constant	0.397**	0.491**	0.278
	(0.13)	(0.18)	(0.20)
BIC	959.4	531.6	430.4
Observations	132	68	64

* p<0.05, ** p<0.01, *** p<0.001; beta coefficients, se in parentheses

Table 5 presents the results of regression analysis of performance on loss aversion and other covariates. Since loss aversion may impact on the performance by encouraging test takers skip more questions, the sign and significance of coefficient estimates of independent variables at Model (2) are similar to Model (1). Thus, the impact of loss aversion and its interaction term is insignificant and I cannot reject the null hypothesis of no effect here as well. The impact of penalty is significant and large because of the point deduction. When I use adjusted scores instead, penalty also becomes insignificant which implies that penalty on its own does not have any other affect other than point deduction.

Table 5. The Impact of the Loss Aversion on Performance under the Penalty System

Dep Var: Score	All b/se	Males b/se	Females b/se
Loss Averse	-2.113	-0.204	-3.769
	(2.45)	(3.62)	(3.26)
Loss Averse x Treatment	-2.009	-7.400	3.626
	(4.98)	(7.24)	(6.74)
Treatment	-10.880**	-11.017*	-10.731*
	(3.84)	(5.58)	(5.19)
Second Session Dummy	-5.768*	-7.529	-3.779
	(2.59)	(3.83)	(3.42)
constant	50.237***	52.214***	48.010***
	(1.92)	(2.79)	(2.59)
BIC	1055.349	556.640	508.760

* p<0.05, ** p<0.01, *** p<0.001; beta coefficients, se in parentheses

5.3. THE GENDER RELATED DIFFERENCES IN GUESSING BEHAVIOUR AND PERFORMANCE UNDER THE PENALTY SYSTEM (HYPOTHESIS 3-4)

The results in Table 6 indicate that there is no gender related impact under the penalty system which is consistent with the previous results where loss aversion is found to have no effect for both genders. The findings, however, do not reject any other variable that might have affected guessing behaviour of participants under the penalty system. My previous and current results imply that any factor that affect the decision making of participants in my sample does not have a gender specific effect and loss aversion has no effect at all.

Table 6. Gender Based Differences in Performance and Guessing

Dep. Var:	Skipped Questions	Score
	b/se	b/se
Female	-0.037	-4.754
	(0.18)	(2.42)
Female x Penalty	-0.303	8.099
	(0.21)	(4.93)
Penalty	3.021***	-15.701***
	(0.22)	(3.80)
Second Session Dummy	-0.562*	-5.728*
	(0.22)	(2.56)
constant	0.350*	51.517***
	(0.14)	(1.90)
BIC	947.676	1052.498

* p<0.05, ** p<0.01, *** p<0.001; beta coefficients, se in parentheses

5.4. INDIVIDUAL INTERVIEW RESULTS

My findings demonstrate that the penalty system encourages skipping more questions; however, there is no systematic discrimination based on different loss aversion levels or gender. These results stand in contrast with the previous literature which at least demonstrates that women skip more under the penalty system than men. I interviewed some of the participants which I believe possess valuable information to shed some light on my findings as well as the overall theory.

Among these participants, there are people who skipped either no or few questions despite their lottery choices indicated the highest level of loss aversion. They explained their behavior with the stakes behind the competition which did not put a high pressure on them as it would be in university entrance exams. Some of them added that they barely knew the answer to many questions which provided them with two choices: either risk answering all questions or skip all the questions they do not know. Since the latter guarantees the defeat, the only option remains to answer all the questions. These people also supported my claim that they wouldn't risk answering all the questions in a university entrance exam.

Those participants who skipped in penalty and/or no penalty system explained their behavior with their affinity towards the questions. They mentioned that they guessed the questions only if they had any idea or had some information about the question which may or may not help omit one of the options and disliked making a pure random choice. It suggests that decreasing number of options from five to four does not necessarily encourage people to guess as disutility from making a pure random choice outweighs positive expected utility from selecting a correct answer after such decision.

CHAPTER 6: CONCLUSION

This paper demonstrates that individuals do not necessarily follow their odds (perhaps even do not calculate it) in deciding to skip or guess a question and they could skip a question under no penalty system, too. Since randomly guessing a question yields positive expected score regardless of the presence or absence of the penalty system in this experiment's settings, skipping a question is not an optimal choice from this perspective and negatively affects the performance as it is further confirmed by the regression analysis. My results, however, do not provide evidence for loss aversion or gender to be a decisive factor for differences in guessing behaviour. It is likely that the lack of high stakes in the knowledge competition or lack of exam pressure in an online format is responsible for these results. I could say with more confidence now that an offline field experiment is more of a necessity for this type of analysis.

The experiment design for the remaining part did work well, and at least could provide consistent results across the most models. The issue is more about the representatives of these results for university entrance or recruitment exams I try to analyse. In order to handle this issue for future research, I would recommend using one of the trial exams for university applicants at a tutoring company instead of a knowledge competition and apply the same randomisation procedure. Since students participate in almost all the trials thanks to family pressure, dropouts from the trial exam are less common and applying randomisation procedure (including the measurement of the loss aversion level) before the exam should not bring negative consequences. I would do the whole process in one session only since the sample would be very large and randomisation at these levels should also handle endogeneity with respect to knowledge level. The biggest challenge for high school students would be to calculate their loss aversion levels. Although some readers may find the current test simple enough, I believe a simpler and more accurate test is needed.

APPENDIX A: LOSS AVERSION SURVEY

You are given 10 000 HUF for participating in this game. The game has two stages. In the first stage, you take one or more of the provided lottery tickets. In the second stage, we will randomly select one of the lotteries, if you have the ticket for the selected lottery, then we will play that lottery. If you don't have the ticket, we don't play any lottery. More lottery tickets you take, higher the chance you will play the lottery. If no lottery is played, you will receive no extra money, but you will keep the initial amount.

How to win the lottery: The chances of winning/losing lottery is 50%/50%. We will toss a coin and if it is Heads, you will gain additional 10 000 HUF in the lottery, if it is Tails, then you will lose the specified amount below. If you end up losing money, we will subtract that amount from initial 10 000 HUF you have received in the beginning of the game.

Lottery	Payoff if you win	Payoff if you lose
A	+ 10 000 HUF	– 2500 HUF
B	+ 10 000 HUF	– 5000 HUF
C	+ 10 000 HUF	– 7500 HUF
D	+ 10 000 HUF	– 10 000 HUF

Now we are in the first stage. Which lottery tickets would you like to take? Select one of the options below. (Note: Watch the tutorial if the game is not clear >> https://youtu.be/HgFMoSOenuo)

- A,B,C,D
- A,B,C
- A,B
- A

REFERENCES

Akyol, Ş. P., Key, J., & Krishna, K. (2019). *Hit or miss? test taking behavior in multiple choice exams* (No. w22401). National Bureau of Economic Research.

Baldiga, K. (2013). Gender differences in willingness to guess. *Management Science*, *60*(2), 434-448.

Ben-Shakhar, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: the role of differential guessing tendencies. Journal of Educational Measurement, 28(1), 23-35.

Burton, N. W., Lewis, C., & Robertson, N. (1988). Sex differences in SAT® scores. ETS Research Report Series, 1988(2), i-23.

Elsesser, K. (2016, July 1). Is The College Board Making The SAT More Difficult For Women? Retrieved from Forbes: https://www.forbes.com/sites/kimelsesser/2016/07/01/is-the-college-board-making-the-sat-more-difficult-for-women/#1c19c4a15f3

Harbaugh, W. T., Krause, K., & Vesterlund, L. (2002). Risk attitudes of children and adults: Choices over small and large probability gains and losses. Experimental Economics, 5(1), 53-84.

Karle, H., Engelmann, D., & Peitz, M. (2019). Student Performance and Loss Aversion.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. Econometrica, 47, 263-291.

Krawczyk, M. (2011). Framing in the field. A simple experiment on the reflection effect. University of Warsaw Faculty of Economic Science Working Paper, (14), 54.

Moore, E., & Eckel, C. (2003). Measuring ambiguity aversion. Unpublished manuscript, Department of Economics, Virginia Tech.

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

Ramos, I., & Lambating, J. (1996). Gender differences in risk-taking behavior and their relationship to sat-mathematics performance. School Science and Mathematics, 96(4), 202-207.

State Examination Center of the Republic of Azerbaijan (2018). The statistical analysis of the exam results. Abituriyent 12 journal

Schubert, R., Brown, M., Gysler, M., & Brachinger, H. W. (1999). Financial decision-making: are women really more risk-averse?. American economic review, 89(2), 381-385.